Masters Program in Geospatial Technologies

ANALYSIS OF PANORAMIO PHOTO TAGS IN ORDER TO EXTRACT LAND USE INFORMATION

Milan Šećerov

Dissertation submitted in partial fulfilment of the requirements for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School

Universitat Jaume I

Westfälische Wilhelms-Universität Münster

ANALYSIS OF PANORAMIO PHOTO TAGS IN ORDER TO EXTRACT LAND USE INFORMATION

Dissertation supervised by:

Prof. Doutor Marco Painho

NOVA Information Management School (NOVA IMS),

Universidade Nova de Lisboa, Lisbon, Portugal.

Dissertation Co-supervised by:

Jacinto Estima, PhD Student

NOVA Information Management School (NOVA IMS),

Universidade Nova de Lisboa, Lisbon, Portugal.

Prof. Doutor Sven Casteleyn

Department of Mathematics,

Universitat Jaume I (UJI), Castellon, Spain.

February 2015

ACKNOWLEDGMENTS

First of all, I want to thank my family. They were always near me, in good and bad moments. So, it was the case with these studies also. They always believed in me even if I didn't have self-confidence in some moments. Without them I wouldn't be able to complete these studies nor this thesis.

Right after this, I want to thank Prof. Dr. Marco Painho (NOVA IMS), my professor and my supervisor, who was always willing to give me right advices and guidelines about the thesis and during my studies in general. He is not only a great professor, but also a great man. I want to thank Jacinto Estima (NOVA IMS), my co-supervisor, for giving me valuable and altruistic advices during the whole thesis progress. I also want to thank Prof. Dr. Sven Casteleyn (UJI) for being my co-supervisor.

My special thanks go to European Commission for funding my studies. Without the financial support, I would never be able to study outside of my country, which is in a very hard situation, nor to find appropriate job at home. In this manner, I want to thank Prof. Dr. Christoph Brox (IFGI) as coordinator of this programme.

I would like to thank Ivo Figueira, my colleague and friend, who helped me to a large extent with doing my thesis. Without his help, it would be much harder to figure out everything about software that I used and that he is familiar with.

At the end, I want to thank to all my friends and colleagues, those from my country and those who are here, for giving me enormous support in hard moments and celebrating with me in good moments.

Thank You Lord for shaping my path in this way!

ANALYSIS OF PANORAMIO PHOTO TAGS IN ORDER TO EXTRACT LAND USE INFORMATION

ABSTRACT

In the recent past, hardly anyone could predict this course of GIS development. GIS is moving from desktop to cloud. Web 2.0 enabled people to input data into web. These data are becoming increasingly geolocated. Big amounts of data formed something that is called "Big Data". Scientists still don't know how to deal with it completely. Different Data Mining tools are used for trying to extract some useful information from this Big Data. In our study, we also deal with one part of these data - User Generated Geographic Content (UGGC). The Panoramio initiative allows people to upload photos and describe them with tags. These photos are geolocated, which means that they have exact location on the Earth's surface according to a certain spatial reference system. By using Data Mining tools, we are trying to answer if it is possible to extract land use information from Panoramio photo tags. Also, we tried to answer to what extent this information could be accurate. At the end, we compared different Data Mining methods in order to distinguish which one has the most suited performances for this kind of data, which is text. Our answers are quite encouraging. With more than 70% of accuracy, we proved that extracting land use information is possible to some extent. Also, we found Memory Based Reasoning (MBR) method the most suitable method for this kind of data in all cases.

KEYWORDS

User Generated Geographic Content

Geographic Information Systems

Data Mining

Predictive Modeling

Panoramio

Photos

Tags

Land use/Land cover

ACRONYMS

API - Application User Interface

CBT - Computer-Based Training

CLC - Corine Land Cover

CORINE - Coordination of Information on the Environment

EEA - European Environment Agency

EIS - Enterprise Information Systems

ETRS89 - European Terrestrial Reference System 1989

GI - Geoinformation

GIS - Geographic Information Systems

HTML - HyperText Markup Language

LULC - Land Use/Land Cover

MBR - Memory Based Reasoning

MMU - Minimum Mapping Unit

NRC - National Reference Centres

OLAP - Online Analytical Processing

OSM - OpenStreetMap

SAS - Statistical Analysis System

SVD - Singular Value Decomposition

UGC - User Generated Content

UGGC - User Generated Geographic Content

UTM - Universal Transverse Mercator

VGI - Volunteered Geographic Information

WGS84 - World Geodetic System 1984

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

CHAPTER ONE

1. INTRODUCTION

1. 1. Background of the Study

In the near past, producing geoinformation (GI) was exclusively done by highly skilled and trained people (Goodchild and Glennon, 2010). This was a very expensive way of producing GI (Goodchild, 2008). With the appearance of the Web 2.0, new possibilities became available. People could not only read and surf through the Web, but also contribute with their own data (Elwood, Goodchild and Sui, 2012). One idea was born - allowing people to insert geographic information from their own perspective (Elwood et al., 2012). This is called VGI (Volunteered Geographic Information) (Goodchild, 2007). These inputs are not done by highly skilled people. Another terms for this are Neogeography (Turner, 2006), Naive geography (Egenhofer, Max and Mark, 1995), Crowd-sourcing geospatial data (Hudson-Smith, Batty, Crooks and Milton, 2009), and they are all related with a type of User Generated Content (UGC) (Goodchild, 2007), etc. Goodchild (2007) proposed the term Volunteered Geographic Information (VGI) to describe the actions of thousands of individuals who are now contributing User-Generated Geographic Content (UGGC) to the Web (Goodchild, 2008). There are now literally hundreds of Web services that collect, compile, index, and distribute VGI content. Wikimapia encourages users to "describe the whole world", OpenStreetMap is developing a free digital map of the world, and Flickr is compiling a vast resource of georeferenced photographs (Goodchild, 2008). Terms like VGI and UGGC are generally used to describe similar variations of Geographic Information (GI) (Spyratos and Lutz, 2014). The main distinction between VGI and UGGC is that people participating in making VGI do it consciously, but in the case of UGGC they mostly do it for fun. People's expressions, which people enter into different websites as UGGC, are called folksonomy (Peters, 2009). This word is similar to taxonomy, but the difference is that taxonomy is structured and folksonomy seems like being chaotic. There are some efforts in trying to make meaningful taxonomy out of folksonomy.

Social media gave opportunity for people to communicate, share their life events, to input more and more data. In the Web, there is huge amount of unsorted data. This represents one part of what is called Big Data (Manyika et al., 2011). Data mining is one scientific field, which is trying to extract some useful information from Big Data as well as other kinds of data. Initiatives like Flickr (http://www.flickr.com) and Panoramio

(http://www.panoramio.com/) enabled the uploading of photos that are geotagged (they have a location) with their metadata descriptions (tags), name of photo, name of author, date, etc (Estima and Painho, 2014).

Scientists are producing LULC (Land Use/Land Cover) maps of the Earth, which are representing respectively human and natural phenomena in respect to the land (Cihlar and Jansen, 2001). Combining UGGC with real scientific information seems attractive and it can bring more information or serve as validation of scientific work.

Trying to extract useful information about LULC from photos' tags has become a topic of research. Is it possible to extract useful information about LULC of certain spot by using UGGC? There are efforts on answering this question (Estima and Painho, 2014; Estima, Fonte and Painho, 2014). One way of answering it is using Data Mining tools (Hughes, O'Connor and Jones, 2012).

In general, if science and technology succeed in managing UGGC, it could be an enormous contribution to society. In this case, even a small contribution of every single human on Earth could be enormous if it is collected and sorted in a meaningful way.

1. 2. Statement of the Problem

The problem which this study is trying to solve is discovering if it is possible to extract useful information about land cover by using UGGC, more precisely, folksonomy. Photos from Panoramio initiative are geotagged and tagged also. They are tagged with vernacular language of crowd. It would be great benefit if it would be discovered that this vernacular language can be useful and used to extract valuable information.

The question is also which level of accuracy, obtained from this folksonomy, is needed in order to prove that photo tags are a useful source of information. The question is which level of accuracy is needed in order to have a strong indication that photo tags can be related with land cover type on the certain spots.

This study is using Data Mining methods to answer the previous questions. There is also a problem in determining which Data Mining method should be used in dealing with this kind of data. Defining the appropriate method is also one part of the problem, which is tried to be solved in this study.

1. 3. Objective of the Study

The main objective of this study is to answer the question of whether it is possible or not to extract information about LULC from Panoramio photo tags.

The specific objectives of this study are:

- Achieve the main objective by using Data Mining tools;
- Understanding the influence of familiarity with the language of examining crowd's expressions;
- Compare the performance of different Data Mining methods in analysing this type of data;
- Performing the accuracy assessment of different Data Mining methods and determining to what extent is extracting LULC information possible;
- Building a predictive model which is functional for this type of data;
- Making a contribution in managing UGGC, social media and folksonomy.

1. 4. Research Questions

Based on the type of data and tools used, this study is trying to answer the following research questions:

1. Is extracting land use information from Panoramio photo tags possible?
2. If yes, to what extent is this information accurate?
3. Which Data Mining method is the most suitable for this type of data?

1. 5. Significance of the Study

UGGC could become an enormous source of information. Since Web 2.0 appeared, we are witnesses of enormous amounts of data being stored in the cloud. Managing this Big Data is a huge challenge for science and technology. Being able to classify, sort and manage these data could provide great benefits and unsuspected amounts of information. This study is focused on geographic information, but it is also using data from the cloud. It is using folksonomy and social media initiatives. Could we extract some useful information from data people entered mostly for fun?

The results of this study could make a contribution to this question. At least an indication could be provided and support further research in this area. Also, these results can indicate which Data Mining method is the best suited for this type of data, which is text

(folksonomy). In a broader sense, this study can provide a small, but not insignificant, contribution for further research in UGGC, social media and folksonomy.

1. 6. Structure of the Thesis

The first chapter introduces the background of the study and the statement of the problem, study objectives, research questions and the significance of the study. It also contains the structure and organisation of the thesis. The next part, chapter two, concentrates on a theoretical literature review and related work for this study. This section presents a brief understanding of UGGC, folksonomy, User Generated Content and previous studies performed on data based on these phenomena in order to extract some useful information or to interrogate the suitability of these data for extracting useful information. The third chapter focuses on the general methodology followed, the data sets used in this study and a description of the study areas. This chapter describes all the procedures and techniques applied in this study in order to answer the question whether the extracting of land cover type is possible or not. It also describes procedures and techniques to calculate the accuracy of the information that was obtained. The results and discussions are presented in the fourth chapter. The results are discussed in a comparative and analytical manner. It shows answers to the research questions - about potential of crowd-sourced data, accuracy maintained and determining the best suited Data Mining method for this type of data. The last chapter presents the conclusions and recommendations. In this section, key findings and critical points that need further treatment are presented and highlighted as a recommendation for future work.

CHAPTER TWO

2. LITERATURE REVIEW

2. 1. Definitions

The notion of VGI (Volunteered Geography Information) was introduced in (Goodchild, 2007). It is part of Naive Geography (Egenhofer and Mark, 1995). At the time when Goodchild wrote his article, there was an explosion of interest in using the Web to create, assemble, and disseminate geographic information provided voluntarily by individuals (Goodchild, 2007). Different sites were encouraging people to provide geographic information. Goodchild examined this phenomenon, and tried to answer questions like: what drives people to do this, how accurate are the results, will they threaten individual privacy, how can they augment more conventional sources. He also examined the role of the amateur in geographic observation (Goodchild, 2007). He concluded that VGI has the potential to be a significant source of geographers' understanding of the surface of the Earth. It is a very cheap, but not so reliable source of information. Also, VGI may offer the most interesting, lasting, and compelling value to geographers (Goodchild, 2007).

Egenhofer and Mark defined the notion and concepts of Naive Geography (Egenhofer and Mark, 1995), the field of study that is concerned with formal models of the common-sense geographic world. Naive Geography corresponds to the body of knowledge that people have about the surrounding geographic world. VGI is one part of it. Naive Geography is expected to provide the basis for designing future GIS. There is a need for a link between people's perception of geographic space and its incorporation into software systems (Egenhofer and Mark, 1995). Common-sense reasoning is difficult, and if scientists manage to formalise such data, it would provide excellent results.

Geographic Information Systems (GIS) are rapidly becoming part of the mass media (Sui and Goodchild, 2011). Remarkable conceptual and technological advances in GIS have been made during the 21st century. 'GIS and media', the speculations of Sui and Goodchild more than 10 years ago, became true, and not only this, but also the growing convergence of GIS and social media. This convergence will continue to transform GIS in fundamental ways (Sui and Goodchild, 2011). These scientists believe that the future development of GIS will be on multiple tracks, like developing GeoWeb, Digital Earth, CyberGIS, virtual geographic environments, and cloud computing. Social media are becoming more location aware and people's experience with their environment is more familiar. Input of different spatial data in

social media is becoming an immense source of information. Sui and Goodchild (2011) believe in the possibility of harvesting fruitful research results that are intellectually exciting, technologically sophisticated, and socially relevant.

These data could be used for different purposes. One example is (Goodchild and Glennon, 2010) where they try to use crowd-sourced geographic information in disaster response. This is just one example how VGI i.e. crowd-sourced data can be useful. Geographic Information created by amateur citizens, VGI, has recently provided an interesting alternative to traditional authoritative information from mapping agencies and corporations (Goodchild and Glennon, 2010), although the quality of data is the major concern. The risk of using VGI in emergency situations is often outweighed by the benefits of its use. "Agencies are inevitably stretched thin during an emergency, especially one that threatens a large community with loss of life and property. Agencies have limited staff, and limited ability to acquire and synthesize the geographic information that is vital to effective response. On the other hand, the average citizen is equipped with powers of observation, and is now empowered with the ability to georegister those observations, to transmit them through the Internet, and to synthesize them into readily understood maps and status reports (Goodcild and Glennon, 2010)." The data quality problem needs further research. There is a need for establishing appropriate mechanisms and institutions for building trust in volunteer sources (Goodchild and Glennon, 2010).

Goodchild (2007) proposed the term volunteered geographic information (VGI) to describe the actions of thousands of individuals. New coined term is User-Generated Geographic Content (UGGC) (Goodchild, 2008). Terms like VGI and UGGC are generally used to describe similar variations of Geographic Information (GI) (Spyratos and Lutz, 2014). The main distinction between VGI and UGGC is that VGI contributors are doing this consciously, but people make UGGC mostly for fun. There are now hundreds or thousands of Web services that collect, compile, index, and distribute VGI content. Wikimapia encourages users to “describe the whole world”, OpenStreetMap is developing a free digital map of the world, and Flickr is compiling a vast resource of georeferenced photographs (Goodchild, 2008).

2. 2. Folksonomy, Tagging and Photo Sharing Sites

How Flickr helps us make sense of the world? Flickr is a very similar initiative to Panoramio. The advent of media-sharing sites has drastically increased the volume of community-contributed multimedia resources available on the web (Kennedy et al., 2007).

These collections generated new opportunities, and new challenges, to multimedia research. Flickr (just like Panoramio) supports photo, time, descriptions and location metadata. Work done by Kennedy et al. (2007), tried an approach of generating aggregate knowledge in the form of "representative tags" for arbitrary areas, and used a tag-driven approach to automatically extract place and event semantics from Flickr tags, based on metadata patterns (Kennedy et al., 2007). With these patterns, vision algorithms could be employed with greater precision. The authors demonstrated a location-tag-vision-based approach to retrieve images of geography-related landmarks and features from the Flickr dataset (Kennedy et al., 2007). The results suggest that community-contributed media and annotation can improve and enhance our understanding of the world (Kennedy et al., 2007).

A similar topic can be found in (Newsam, 2010). With a very similar introduction to the paper above (Kennedy et al., 2007), the author is leading us into the Flickr based possibilities of extracting useful information. His focus is on learning the correspondence between the textual tags and the visual content. This article focuses on knowledge discovery based on the geographic location of social media. The data used are large collections of georeferenced community-contributed photographs, such as those available on Flickr or Panoramio. The primary goal was to connect this research thrust to the larger phenomenon of VGI. In particular, the author argued that georeferenced social media is another form of VGI and the geographic discovery it enables is in effect crowd-sourcing what is where on the Earth's surface.

This all wouldn't be possible if there is no presence of a tagging system. "In recent years, *tagging systems* have become increasingly popular. These systems enable users to add keywords (i.e., “tags”) to Internet resources (e.g., web pages, images, videos) without relying on a controlled vocabulary. Tagging systems have the potential to improve search, spam detection, reputation systems, and personal organization while introducing new modalities of social communication and opportunities for Data Mining. This potential is largely due to the social structure that underlies many of the current systems. (Marlow et al., 2006)" Despite the rapid expansion of applications that support tagging systems, it is yet not well studied or understood (Marlow et al., 2006). The author proposes a tagging system which would have its taxonomy and be more formal. He also provides a short description of related academic work. There is still a problem that ordinary people are not scientists. If this would be the case, we wouldn't have a problem with UGGC.

Specia and Motta (2007) provide a more realistic way of thinking. They say: "...the use of the same tags by more than one individual can yield a collective classification schema."

They present an approach for making explicit the semantics behind the tag space in social tagging systems, so that this collaborative organization can emerge in the form of groups of concepts and partial ontologies. They achieved this by using a combination of shallow pre-processing strategies and statistical techniques together with knowledge provided by ontologies available on the semantic web. Their preliminary results on the del.icio.us and Flickr tag sets show that their approach is very promising. It generates clusters with highly related tags corresponding to concepts in ontologies and meaningful relationships among subsets of these tags can be identified (Specia and Motta, 2007).

2. 3. Previous Research Works on Photo Sharing Sites

There are many papers dealing with Flickr and Panoramio photos and their metadata. Positional accuracy analysis of Flickr and Panoramio photos done by (Zielstra and Hochmair, 2013) shows some issues that could be an obstacle for using this source of data for extracting some useful information. This study analyses the positional accuracy of 1433 photos from 45 areas by comparing the geottaged position of photos to the manually corrected camera position based on the image content (Zielstra and Hochmair, 2013). The authors came to the clue that Panoramio photos have better positional accuracy than those from Flickr. This is because Panoramio is considered as a more serious instance, while in the case of Flickr, people are sharing everything without too much caring of tagging and geotagging photos. Also, in different world regions, the positional accuracy is different. This is the case with the image category, too (Zielstra and Hochmair, 2013). Authors are concluding that these findings can be helpful when considering Flickr and Panoramio images as data sources for future geo-applications and services.

Many papers present different ways of using photo sharing sites in obtaining different types of information. One of them (Hollenstein and Purves, 2014) is dealing with exploring place through user-generated content by using Flickr tags to describe city cores. This study describes how everyday or vernacular language terms are used to describe certain areas in a city. They explored such language by harvesting georeferenced and tagged metadata associated with 8 million Flickr photos. Using Flickr metadata, it is possible not only to describe the use of a term, but also to explore the borders of different city areas at the level of individual cities, whilst accounting for bias by the use of tag profiles (Hollenstein and Purves, 2014). This paper starts by setting out a number of examples of the uses of vernacular geography in examples of putative information systems, and argued that such geographies were not often captured by current administrative representations. The authors also argued that user generated content from sources such as Flickr might provide one way

of exploring such vernacular geography, and in particular both specific and generic use of place names in urban areas where Flickr predominates (Hollenstein and Purves, 2014).

In another, but similar paper, the authors are trying to perform event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections (Kisilevich et al., 2010). The authors argue that millions of geotagged photos pose new challenges in the domain of spatio-temporal analysis. In this paper, several different tasks are defined related to analysis of attractive places, points of interest and comparison of behavioral patterns of different user communities on geotagged photo data. They performed the analysis and comparison of temporal events, rankings of sightseeing places in a city, and they studied the mobility of people using geotagged photos (Kisilevich et al., 2010). The authors took a systematic approach to accomplish these tasks by applying scalable computational techniques, using statistical and data mining algorithms, combined with interactive geo-visualization. They provided an exploratory visual analysis environment, which allows the analyst to detect spatial and temporal patterns and extract additional knowledge from large geotagged photo collections. The authors also demonstrated their approach by applying the methods to several regions in the world (Kisilevich et al., 2010). They analysed the structure of the event-based movement data in order to define systematically several tasks for event-based analyses of people's travel activities, behavior and mobility using geotagged photo data, collected and shared by people from all over the world (Kisilevich et al., 2010). They concluded that attractiveness of places can be obtained from photo data. A way of assessing attractiveness of places based on these data is also presented. In this complex study, it is shown how photo sharing sites can be useful when explored in a good way.

There is another interesting approach to this kind of data. It consists of (Leung and Newsam, 2010) inferring what-is-where from georeferenced photo collections. They used what they called "proximate sensing". "The primary and novel contribution of this work is the conjecture that large collections of georeferenced photo collections can be used to derive maps of what-is-where on the surface of the earth. (Leung and Newsam, 2010)" They investigated the application of what they termed “proximate sensing” to the problem of land cover classification for a large geographic region. They showed that their approach is able to achieve almost 75% classification accuracy in a binary land cover labeling problem using images from a photo sharing site in a completely automated fashion. They investigated how existing geographic knowledge can be used to provide labeled training data in a weakly-supervised manner. Also, they investigated the effect of the photographer’s intent when he or

she captures the photograph, and a method for filtering out non-informative images (Leung and Newsam, 2010). Their results are quite interesting. They came to a clue that weakly-supervised labeled training data resulted in better performance than manually labeled training data. They found out that photographer's intention is very important in the relevance of a photo. They also tried to filter images with faces, but it didn't yield better results (Leung and Newsam, 2010).

Wang, Korayem and Crandall (2013) also presented an interesting way of using photo sharing sites. They used billions of public photos to investigate latent visual information about the world. In this case, they were trying to recognise snowy areas. They studied the feasibility of observing the state of the natural world by recognising specific types of scenes and objects in image collections. They tried to recreate satellite maps of snowfall by automatically recognising snowy scenes in geotagged and timestamped images (Wang, Korayem and Crandall, 2013). Their best result, using modern vision techniques, achieved 81% of accuracy.

2. 4. Works on Examining the Potential of the Photo Sharing Sites in Extracting the Information about Land Use

The European Environment Agency (EEA) published "CLC 2006 technical guidelines". This document is a guideline for understanding the Corine Land Cover map from 2006, the CORINE programme, organisation of the project, scales used, etc.

The minimum mapping unit used in this programme has size of 25 hectares. CORINE land cover nomenclature defines 3 hierarchical levels of land cover types. In the first level there are 5 types. In the second level these 5 types are split in 15 types. In the third level these 15 types are split in 44 land cover types. The mapping scale chosen for the project was 1:100.000 (for the third level). For this study, only the first level of land cover types will be used:

1. Artificial surfaces;
2. Agricultural areas;
3. Forests and semi-natural areas;
4. Wetlands;
5. Water bodies.

In this guideline methods used, preliminary work, collection and organisation of data, procedure and field work are also described.

All papers presented in this section describe improvements to land cover by examining the potential and suitability of geotagged images for this purpose. Exploring geotagged images for land use classification in (Leung and Newsam, 2012) is an example of investigating the problem of geographic discovery, particularly land use classification, through crowd-sourcing of geographic information from geotagged photo collections. Their results show that the visual information contained in these photo collections enables the extraction of three classes of land use on two university campuses. They found out that text entries accompanying these photos are informative for geographic discovery using visual and textual features of photos at both individual and group image level. The results of this approach gave promising first steps on this interesting but challenging problem (Leung and Newsam, 2012).

There is quite a lot of research dealing with exploratory analysis of UGGC for land use classification. Estima and Painho (2013) presented exploratory analysis of OpenStreetMap for land use classification. They used the Corine Land Cover database as reference and continental Portugal as study area. They developed a comparative analysis of OpenStreetMap (OSM) and Corine Land Cover and evaluated the quality of OSM polygon features classification from the first level of the nomenclature. They also analysed the spatial distribution of OSM classes over continental Portugal and obtained 76% of accuracy for global classification. In a later work of the same authors (Estima and Painho, 2015), they reviewed the existing literature on using OSM data for LULC database production and moved this research forward by exploring suitability of the OSM Points of Interest dataset. They concluded that OSM can give very interesting contributions and that OSM Points of Interest dataset is more suitable for those areas classified as artificial surfaces. In a previous work of Estima (2012), "Using Volunteered Geographic Information to help Land Use/Land Cover mapping", the author provides an overview of what might been done in this area.

These two authors also dealt with Flickr geotagged and publicly available photos in order to estimate its adequacy for helping quality control of Corine Land Cover (Estima and Painho, 2013). Preliminary analysis of the adequacy of photos from the Flickr initiative in order to use them as a source of field data in the quality control of the LULC database production is presented in (Estima and Painho, 2013). The authors evaluated its temporal and spatial distribution over Continental Portugal and also its distribution over LULC classes using as a reference the European Corine Land Cover database (Estima and Painho, 2013). Their conclusion is that this source is very valuable but needs to be combined with other sources because of some issues related with spatial distribution (Estima and Painho, 2013). Estima

and Painho (2014) also continued their study about suitability of photo based Volunteered Geographic information initiatives in order to help quality control of Corine Land Cover. "This paper conducts a preliminary analysis of the adequacy of photos from Flickr and Panoramio initiatives in order to use them as a source of field data in the quality control of the Land Use/Cover classes using a reference the European Corine Land Cover database (Estima and Painho, 2014)." The conclusion was that this source is very valuable but needs to be combined with other sources due to its uneven spatial distribution. (Estima and Painho, 2014). More photos are present in the cities than in surrounding areas. This is more common for photos coming from Flickr initiative. Also, artificial surfaces is a class with much more photos available than in other classes. Authors truly believe that this is due to tourism based photos. Regarding time, there are more photos in summer period (Estima and Painho, 2014).

Going more further, comparative study of Land Use/Cover classification was performed using Flickr photos, satellite imagery and Corine Land Cover database (Estima, Fonte and Painho, 2014). This is one attempt of evaluation of geo-referenced and publicly available photos from the Flickr initiative. The question was whether publicly available photos from the Flickr initiative could be used as a source of geographic information to help Land Use/Cover classification (Estima, Fonte and Painho, 2014). The authors compared the classification obtained for selected photo locations against the classification obtained from high resolution satellite imagery for the same locations (Estima, Fonte and Painho, 2014). The conclusion is similar like in their previous works. They concluded that this source cannot be used alone for the purpose of Land Use/Cover classification, but it might be helpful in containing useful information if combined with other sources (Estima, Fonte and Painho, 2014). They identified the problem of taking a photo of one land cover class, but with standing point on another land cover class, like is the case with water bodies when people take pictures of the ocean from the land.

2. 5. Usage of Data Mining with Textual Input Variables

This section presents works on the usage of data mining with textual variables, as photo tags could be considered as text. Photo tags are one type of crowd-sourced information. There are several approaches to analyse them, but for this study, the Data Mining approach is of significant importance.

A novel machine learning based approach to determining the semantic relevance of community contributed photo tags is presented in (Hughes, O'Connor and Jones, 2012). "Current large scale community image retrieval systems typically rely on human annotated

tags which are subjectively assigned and may not provide useful or semantically meaningful labels to the images. (Hughes, O'Connor and Jones, 2012)" The authors described a method to improve text based image retrieval systems by eliminating generic and not relevant tags. Using this feature set, machine learning models are trained to classify the relevance of each tag to its associated image (Hughes, O'Connor and Jones, 2012). The evaluation of this method is based on using human annotated collection of Flickr images. This is one example on how machine learning can be used in dealing with photo tags.

Aggarwal and Zhai (2012) presented very useful guidelines about mining text data. Various methods and its use are provided in this book. After the words of authors: "The problem of text mining has gained increasing attention in recent years because of the large amounts of text data, which are created in a variety of social network, web, and other information-centric applications. Unstructured data is the easiest form of data, which can be created in any application scenario. As a result, there has been a tremendous need to design methods and algorithms which can effectively process a wide variety of text applications. This book provides an overview of the different methods and algorithms, which are common in the text domain, with a particular focus on mining methods (Aggarwal and Zhai, 2012)".

In a paper from Patricia Cerrito (2009), "Predictive modeling in Enterprise Miner versus regression", it has been discussed all available methods in predictive modeling in SAS Enterprise Miner. The adequacy of a model is questioned by its ability to predict rare occurrences. It can be highly accurate, but without predicting rare occurrences. In contrast, predictive modeling in Enterprise Miner was designed to accommodate large samples and rare occurrences as well as providing many measures of model adequacy (Cerrito, 2009). Cerrito poses a question: "What do we mean by "best" model? Answers are different if we use different methods, outcome, etc.

Masand, Linoff and Waltz (1992) describe a method for classifying news stories using Memory Based Reasoning (MBR). Although, this paper is from 1992, it is still useful and relevant. MBR is a k-nearest neighbor method. The authors used an database of about 50,000 stories. These codes are assigned to new, unseen stories with a recall of about 80% and precision of about 70% (Masand, Linoff and Waltz, 1992). The authors believe that this approach is effective in reducing the development time to implement classification systems involving large number of topics for the purpose of classification, message routing, etc (Masand, Linoff and Waltz, 1992). They demonstrated that a relatively simple MBR approach enables news story classification with good recall and precision (Masand, Linoff and Waltz, 1992).

CHAPTER THREE

3. DATA AND METHODOLOGY

3. 1. Description of the Study Area

Three regions have been chosen in this study. These are Cambridgeshire in England, Coimbra district in Portugal and South Bačka district in Serbia (Figure 1).

Cambridgeshire is a county in England. It has total area of 3,389km^2, populated with 806,700 inhabitants (2011 est.). The elevation is going between 2,75 and 146m. The biggest city is Cambridge with 122,700 inhabitants (2011 est.). Cambridge is a middle size university city. Large areas of the county are extremely low-lying (http://en.wikipedia.org/wiki/Cambridgeshire, accessed: December, 2014).

Coimbra district is a district in Portugal, located near Atlantic Ocean. Total area of the district is 3,947km^2, populated with 441,245 inhabitants. The capital city of the district is Coimbra (143,396 inhabitants (2011 census)), which is the oldest university city in Portugal. Elevation is going from 9 to 499m (http://en.wikipedia.org/wiki/Coimbra_District, accessed: December, 2014).

South Bačka district is an administrative region in Serbia. Total area of the region is 4,016km^2, populated by 615,371 inhabitants (2011 census). The capital city of the district is Novi Sad (250,439 (2011 census)), which is also a very old university city (http://en.wikipedia.org/wiki/South_Ba%C4%8Dka_District, accessed: December, 2014).

These three regions are chosen because they are similar by size, they are three different speaking regions and all of them are university cities. There is a difference between water bodies. Cambridgeshire is characterised by narrow rivers and streams, Coimbra district has presence of ocean and small streams, and South Bačka district has Danube River, wide and with lively activities on it. The idea was to examine naive people's expressions of three different languages about photos they are taking and uploading to Panoramio initiative and how it corresponds on ability of extracting land use information from it. One criterion of choosing these three regions was the author's familiarity with English and Serbian language, although not so familiar with Portuguese language, but studying in Portugal.

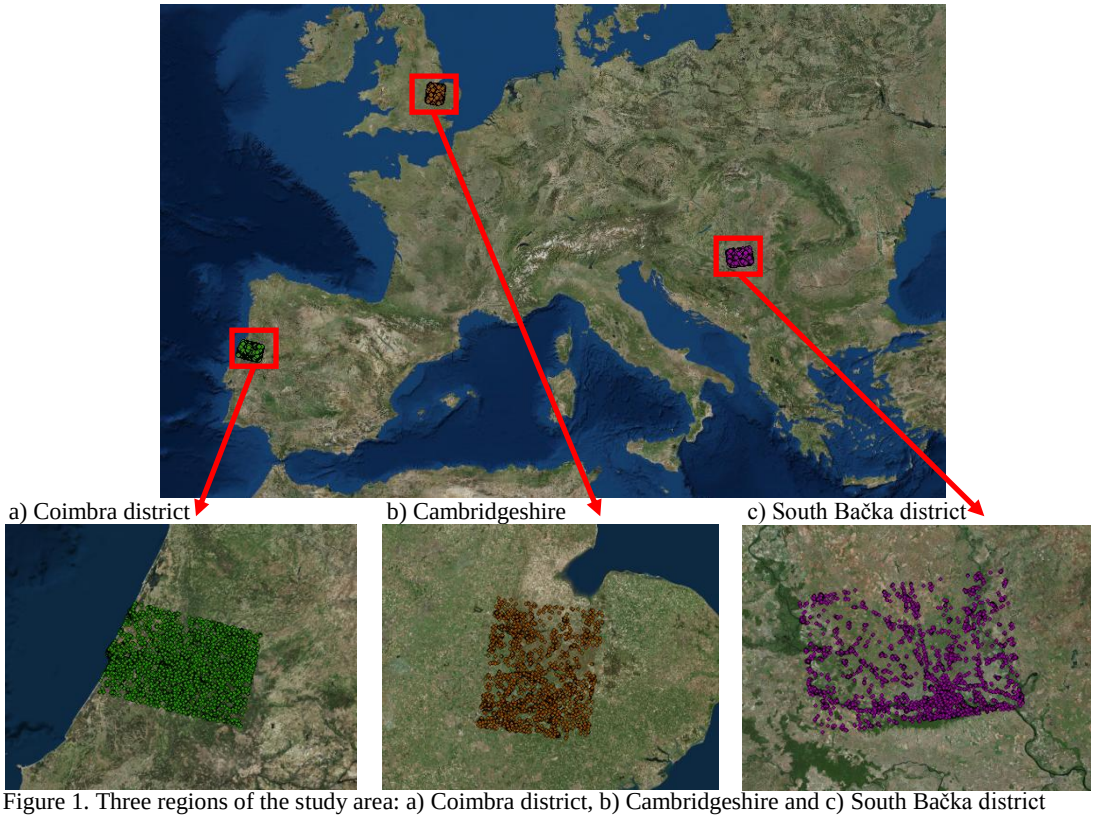

Figure 1. Three regions of the study area: a) Coimbra district, b) Cambridgeshire and c) South Bačka district

Images downloaded from Panoramio website are not exactly inside administrative borders of the mentioned territories because regions from which images were downloaded are defined by bounding boxes and used as the study area like they are (Figure 1). Locations of Panoramio images in green are from Coimbra district. Locations of images shown in brown are from Cambridgeshire. Violet dots represent locations of images in South Bačka district.

Area	SW corner	NE corner	Beginning time	Ending time
Cambridgeshire	x = -0.4999 y = 52.005779	x = 0.51414 y = 52.740341	1128380400 October 4, 2005 00:00h	1412118000 September 30, 2014 24:00h
Coimbra district	x = -8.90902 y = 39.930199	x = -7.73137 y = 40.520390	1128380400 October 4, 2005 00:00h	1412118000 September 30, 2014 24:00h
South Bačka	x = 18.969580 y = 45.152512	x = 20.316111 y = 45.777248	1128380400 October 4, 2005 00:00h	1412118000 September 30, 2014 24:00h

Table 1. Bounding boxes with time period for observed regions

A bounding box is a rectangle defined with the longitude and latitude of the lower left and upper right corners (Hill, 2000). Also, the time period from which images are downloaded should be defined. The time is given like system time, but it can be converted in human friendly time. Bounding boxes of study regions and the time period are shown in Table 1.

3. 2. Data Pre-processing and the Datasets obtained

The procedure for data preprocessing and analysis is depicted in Figure 2. The data were downloaded from Panoramio initiative. It was consisted of images' metadata. These images have a location on the Earth. This location is joined with land use classes obtained from Corine Land Cover database. After this step, the data pre-processing was performed. The data were cleaned and the training, validation and testing datasets were separated. The data analysis was performed by building predictive models using Data Mining and performing accuracy assessment. The pre-processing was performed two times and building the models was repeated as many times as it is concluded that accuracy cannot be improved anymore.

The data is composed by a collection of images' metadata, downloaded from the Panoramio website (http://www.panoramio.com/). Actually, the most interesting things for the study are the tags and coordinates. These images cover the three study areas - Cambridgeshire, Coimbra and South Bačka districts. Also, images are downloaded inside a defined time period. All images have coordinates, which means they are geotagged and they have an exact location on a map.

"Panoramio is a website where users can upload and geolocate photos of the world, explore the world through other people's photos, and join a community of other photography enthusiasts. Geopositioned photos uploaded in Panoramio may be displayed in a Panoramio Group, Google Earth and Google Maps and other sites using the Panoramio API. Depending on the content of the uploaded photo, it will be eligible to be sponsored on the mentioned destinations (http://www.panoramio.com/help/acceptance_policy, accessed: December, 2014)." Panoramio is using WGS84 ellipsoid with modified Universal Transverse Mercator (UTM) conformal projection.

Afterwards, these images (their metadata) are associated with a land cover class which is present on the location of an image. These classes were retrieved from the Corine Land Cover database (http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-3, accessed: September, 2014). In order to perform our study, we needed images' tags (people's descriptions of photos) and land cover classes of the locations where the images were taken. The coordinates of the images served the purpose of associating a specific image with its land cover class at the exact same location. The datasets consist of images' metadata downloaded from Panoramio website and land cover classes associated to them. Associating land cover classes was possible because all of the images are geotagged - they have exact location on Earth's surface. Because different geographic reference systems and projections

are used in Panoramio initiative and CLC, it was necessary to project images' locations to the same projection as CLC uses.

"The Corine Land Cover (CLC) inventory was initiated in 1985 (reference year 1990). Updates have been produced in 2000 and 2006, and the latest 2012 update is under production. It consists of an inventory of land cover in 44 classes. CLC uses a Minimum Mapping Unit (MMU) of 25 ha for areal phenomena and a minimum width of 100 m for linear phenomena. The time series are complemented by change layers, which highlight changes in land cover with an MMU of 5 ha. CLC is produced by visual interpretation of high resolution satellite imagery. The Eionet networks National Reference Centres Land Cover (NRC/LC) are producing the national CLC databases, which are coordinated and integrated by EEA. The 2012 version of CLC is the first one embedding the CLC time series in a structural context of the Copernicus programme http://land.copernicus.eu/pan-european/corine-land-cover, accessed: December, 2014)." The CLC reference system is based upon ETRS89. In our study we used CLC from 2006, but only first level of land cover classes (http://gis.stackexchange.com/questions/62715/corine-land-cover-2000-coordinate-reference-system, accessed: December, 2014).

There are 5 classes in first level of CLC nomenclature. They are:

1. Artificial surfaces,
2. Agricultural areas,
3. Forest and seminatural areas,
4. Wetlands,
5. Water bodies (land.copernicus.eu).

Finally, our datasets consist of people's tags of certain images, IDs of images and land cover class of the point where image is taken (or geotagged). There is a possibility to geotag image after (which is not so reliable) or, if people use smarthphones, images are being geotagged automatically (which is also not extremely accurate - it depends on accuracy of smartphone's GPS system or other locating options (Leung, 2012)).

Initially, there were three datasets, one for each of the three areas. The dataset downloaded for the Cambridgeshire area had 15,237 features (images), the Coimbra dataset had 19,580 features and the South Bačka dataset initially had 20,985 images. But these data exhibited a considerable amount of noise such are images with no tags or not relevant tags, so they were demanding cleaning (pre-processing). After the cleaning of data (2 rounds of cleaning), we finally got three datasets, which were used to form training, validation and testing datasets. These datasets are necessary for data mining analysis and building predictive models.

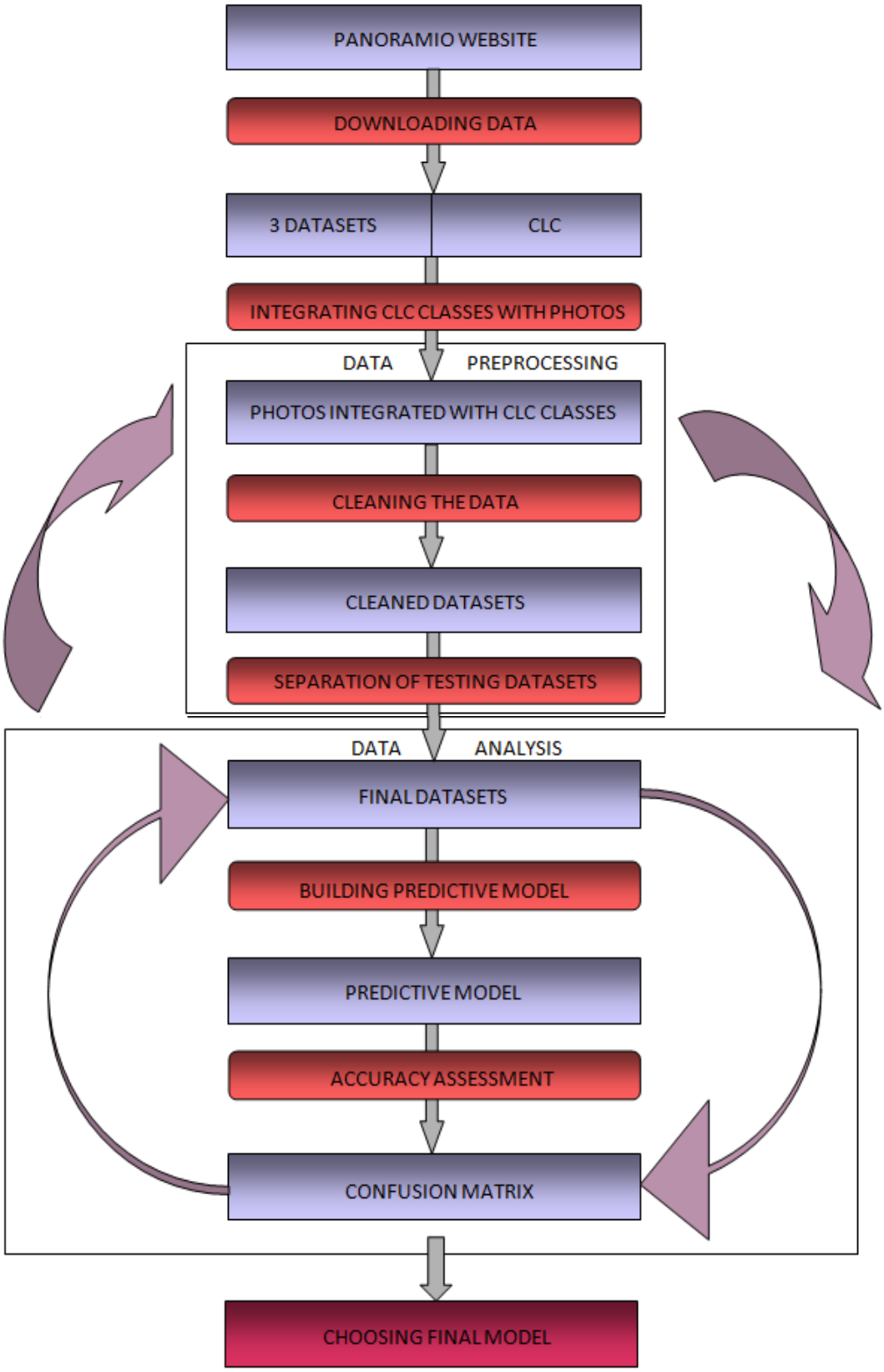

Figure 2. The procedure of the study's data pre-processing and analysis

Final datasets are much smaller than the initial ones:

- The Cambridgeshire dataset has 3,250 images.
- The Coimbra dataset has 3,047 images.
- The South Bačka dataset has 3,790 images.

400 images were randomly selected in each of the datasets and put aside to serve as testing datasets. The rest of images were used for like training and validation procedures. The size of a validation dataset is 20% of initial number of images after deducting the testing dataset.

3. 2. 1. Data Pre-processing

Data were very noisy and demanded cleaning (pre-processing). Data pre-processing or data cleaning was done manually in two rounds. In the first round (described in detail below), data were cleaned roughly and only a small number of the features was deleted. But results were not quite satisfying, so we decided to clean data carefully by separating (erasing) irrelevant data. Irrelevant data means data with tags which are not useful and cannot be correlated with certain land cover class in any way. Here will be shown and explained the first round of cleaning in detail. The second round of cleaning was quite complicated and detailed. In this second round the South Bačka dataset cleaning will be shown like example for other two datasets and to illustrate the complicity of cleaning which was performed.

In the first round of cleaning, which was very general and rough, only a few images were removed comparing to the second round (if we don't count images with no tags, which number was 7,466 for Cambridgeshire, 7,654 for Coimbra district and 9,814 for South Bačka district).

3. 2. 1. 1. Pre-processing of the Cambridgeshire dataset

In Cambridgeshire dataset initially were 15,237 images. After the first round of cleaning 8,224 images were removed of which 758 had single or combined irrelevant tags and 7,466 had no tags. At the end of the first round of data cleaning, 7,013 images remained. Tags which were not relevant for the study are shown in Table 2.

An explanation about reasons of removing mentioned images is at hand. The criterion used to analyse tags and to eliminate images from the original dataset was common sense of the author. For example:

- Images with only dates, numbers, years and symbols (letters which software doesn't recognize) were removed because they are not useful nor can they indicate the existence of a land cover class at this spot.

- "England", "UK", "GB", "Wales" or tags which are names of shires are too general and cannot provide any useful information.
- Names of cities and towns are not relevant because it is possible to take a picture inside the town, at some distance (standing on another land cover class) or we can travel around and taking pictures standing on all land cover classes, but tagging with the name of the town where we originally came to visit or where we live.
- "Snow", "autumn", "sunset", "sky", etc. are notions which can be present at any land cover class. They are very general and they only confuse data mining software during building a model (also tags like "animal", "birds", "trees", etc.). Given names are not useful because it is possible to take a picture of a person at any environment.
- Images with no tags (almost half of them), were also removed.

Tag	Number	Tag	Number	Tag	Number
Numbers	17	Years	39	Datum	17
"England"	128	"UK"	97	Given names	3
"merged"	1	"best"	60	"GB"	9
"Google earth"	11	"Wales"	3	"snow"	11
"Trogir"	1	"3D"	2	Symbols without meaning	47
"Anglia"	13	"Inglaterra"	25	"animals"	48
"autumn"	1	"birds"	3	"Border collies"	1
"Cambridgeshire"	126	"sunset"	2	"sky"	1
"Essex"	5	"Europe"	1	"Huntingtonshire"	13
"Lincolnshire"	3	"MW"	1	"National trust"	22
"Not in Google"	6	"Old England"	23	"trees"	3
"trip"	2	"West Europe"	11	"wolf"	1
Initial number of images				15,237 (100%)	
Removed images with irrelevant tags				758 (4.97%)	
Removed images with no tags				7,466 (49%)	
Total images removed				8,224 (53.97%)	
Total images remained				7,013 (46.02%)	

Table 2. Tags of images in Cambridgeshire dataset that were removed (1st round cleaning)

After building the models with this dataset, it was decided that detailed cleaning should be performed in order to try to improve the results. So, after analysis of these data, the second round of data cleaning was performed. Table is shown in Annex A.
Finally, we had 3,250 images in the Cambridgeshire dataset after the second round of cleaning. This dataset is split in two parts, 2,850 images for training and validation and 400 images for the testing dataset. 11,987 images were removed in total.

3. 2. 1. 2. Pre-processing of the Coimbra district dataset

The Coimbra dataset initially had 19,580 images. After the first round of cleaning, 10,094 images remained. So, 9,486 images were removed of which 7,654 with no tags and 1,832 with single or combined irrelevant tags. Tags which were not relevant for the study are shown in Table 3.

Tag	Number	Tag	Number	Tag	Number
Years	72	Datum	3	"distrito"	34
"Portugal"	656	Numbers	19	"diversos"	17
"Coimbra"	855	"Contest"	5	"Europa"	1
Symbols	10	"best"	64	"nenhum"	63
Given names	3	"HDR"	3	"outros"	8
"Google earth"	1	"by night"	5	"PT"	13
Initial number of images				19,580	(100%)
Removed images with irrelevant tags				1,832	(9.36%)
Removed images with no tags				7,654	(39.09%)
Total images removed				9,486	(48.45%)
Total images remained				10,094	(51.55%)

Table 3. Tags of images in Coimbra dataset that were removed (1st round cleaning)

An explanation about reasons of removing mentioned images is at hand. The criterion used to analyse tags and to eliminate images from the original dataset was common sense of the author. The examples of reasons for eliminating images are:

- Images with only years, dates and numbers are obviously irrelevant for the study. These tags cannot indicate anything about which land cover class is present on a certain location.
- Tags like "Portugal", "Coimbra" or "Europe" are names of geographic regions. They are too general and not useful for the study.

- "Symbols" are words with strange letters, probably because software couldn't recognize them, so, they are also not useful.
- "Diversos", "nenhum" and "outros" are too general terms also. Given names are not useful, as we explained before.
- "By night" could be anything, city, fields, forest, boat trip, etc. So, it is also not useful for the study.

After this first round of the cleaning, we got some results, but in order to improve them, we performed the second detail round of cleaning, which results will be shown in Annex B. This cleaning is performed according to common sense, but with some problems because of not familiarity with Portuguese language. We will see later how this affected our results.

After the second round of cleaning, we had 3,047 images in the Coimbra district dataset. This dataset was split in two parts, 2,647 images for training and validation and 400 images for testing dataset. Finally, 16,533 images have been removed in total.

3. 2. 1. 3. Pre-processing of the South Bačka district dataset

The South Bačka district dataset initially had 20,985 images. After the first round of cleaning, 9,697 images remained. So, 11,288 images were removed of which 9,814 with no tags and 1,474 with single or combined irrelevant tags. Tags which were not relevant for the study are shown in Table 4.

The data were cleaned manually. The criterion used for cleaning was common sense of the author. The reasons for removing images with tags from Table 4 are obvious:

- Tags with only numbers, years and dates cannot indicate any land cover class.
- "Serbia", "Croatia", "Vojvodina" or "Bačka" are regions (or countries) and cannot provide any relevant information.
- Symbols are consequence of different letters used in Serbian language, so software couldn't recognize them. Accordingly, they are also irrelevant.
- "Omiljeni", "razno" and "ostalo" means "favourite", "divesity" and "rest", respectively. These words by itself are too general for this purpose.
- "People", "photo" and "beautiful" are also general terms in this case.
- Abbreviations like "S&MN" and "SRB", which are abbreviations of "Serbia & Montenegro" and "Serbia", are the same like it was the case with the names of regions or countries.

Tag	Number	Tag	Number	Tag	Number
Numbers	15	Years	134	Datum	36
"Serbia"	306	...	4	"Croatia"	46
Symbols	486	"best"	101	"Vojvodina"	195
"beautiful"	2	"contest"	3	Given names	17
"HDR"	12	"Bačka"	32	"Jugoslavija"	3
"makro snimci"	24	"omiljeni"	4	"ostalo"	16
"people"	4	"photo"	1	"razno"	3
"S&MN"	1	"SRB"	2	"srbsko"	27
Initial number of images				20,985 (100%)	
Removed images with irrelevant tags				1,474 (7.02%)	
Removed images with no tags				9,814 (46.77%)	
Total images removed				11,288 (53.79%)	
Total images remained				9,697 (46.21%)	

Table 4. Tags of images in South Bačka dataset that were removed (1st round cleaning)

After the first round of cleaning, the second round was performed in order to try to improve the results that we have got. The table with the removed tags of the second round of cleaning will be provided below (Table 5). The tags were observed carefully and classified like relevant or irrelevant by using common sense.

After the second round of cleaning, we had 3,790 images. These images are divided into two parts, 3,390 images for training and validation and 400 images for testing dataset. In total, we removed 17,195 images from this dataset.

In the Table 5 we are going to show complete and detail tags which are cleaned in South Bačka dataset. The names of settlements and few more tags were removed in the second round of data cleaning. They were found like irrelevant by using common sense.

Tag	Num	Tag	Num	Tag	Num	Tag	Num
Symbols	176	"Serbia"	475	"Croatia"	94	"Novi Sad"	1,173
"Vojvodina"	134	"Apatin"	5	"Bač"	68	"Bečej"	85
"Kula"	104	"Novi Bečej"	73	"N. Miloš."	7	"Panonija"	23
"Vukovar"	411	"Sr. Karl."	89	"Sombor"	170	"Rekovac"	2
"glavna	18	"Petrovaradin"	89	"Kovilj"	53	"Ilok"	46

ruta"							
"Tovarnik"	2	"Budisava"	11	"Sr. Kam."	31	"Banat"	10
"Srbobran"	6	"Ada"	1	"areal shots"	2	"životinje"	60
"Apatin"	13	Datum	8	"autoput"	31	"automobili"	16
"autumn"	38	"Mol"	19	"Odžaci"	69	"B. Topola"	18
"Beočin"	24	"Bogojevo"	4	"Futog"	33	"pčele"	2
"Banoštor"	11	"Begeč"	4	"Belgrad"	1	"best"	183
"B&W"	2	"Bocke"	40	"iz aviona"	10	"beautiful"	8
"Radičević"	23	"Doroslovo"	20	"Gajdobra"	16	"Rakovac"	1
"Mali Iđoš"	77	"Feketić"	213	Given names	55	"Borovo"	19
"železnice"	686	"zima"	8	"sunset"	17	"Neštin"	1
"B.D. Polje"	1	"gasovod"	14	"Šarengrad"	3	"clouds"	7
"Deronje"	6	"Ilok"	20	"contest"	9	"Lovas"	1
"cveće"	14	"Dalj"	153	"Gunaras"	3	"svitanje"	1
"insekti"	19	"Erdut"	2	"Kać"	19	"Koruška"	15
"Lovćenac"	10	"medo"	1	"Titel"	40	"odmor"	3
"Čurug"	6	"Sivac"	5	"Vrbas"	1	"Despotovo"	4
"Temerin"	85	"Krčedin"	2	"Aradac"	7	"B. Petrov."	14
"ljudi"	3	"Ledinci"	11	"Kulpin"	12	"Kucura"	7
"Irig"	7	"Inđija"	5	"by bike"	4	"B. Grad."	35
"B. P. Selo"	5	"Čenej"	12	"Čelarevo"	6	"slikano mobilnim"	2
"leto"	5	"plant"	1	"Čortanovci"	23	"B. Palanka"	60
"Bajša"	1	"Melenci"	20	"Beška"	10	"night"	2
"Bukovac"	6	"Đurđevo"	5	"Bačka"	39	"Kisač"	35
"Begeč"	1	"R. Krstur"	9	"Bođani"	7	"Crvenka"	3
"Slavonija"	4	"Crni Steva Silbaš"	1	"Crni peče meso"	1	Total:	5,907 (60.92%)

Table 5. Tags of images in South Bačka dataset that were removed (2nd round cleaning)

A general explanation of the reasons why certain tags were used to eliminate images follows.

- Most of the tags above are names of smaller towns and villages in South Bačka district. They were standing alone or in combination with country or region name. Anyway, these tags are irrelevant for determining or indicating some land cover class because people tag image with a name of a village, but taking picture all

around village, in forests, fields, inside village. It seems (and it is true) that people didn't really care about tagging images and giving better description.

- Other tags whose translation are like flowers, animals, sunset, sky, winter, vacation, etc. are also not relevant for determining land cover class because they are too general and could be taken in any land cover class.
- At the end, we found really funny tags like is "Crni peče meso", which means "black guy is frying meat". He could fry meat in forest doing barbecue, in his house or on the river boat, so, this tag is, beside it is funny, also irrelevant.

Summary table with the main numbers of images for each of the three sites is provided in Table 6.

SUMMARY TABLE	CAMBRIDGESHIRE	COIMBRA DISTRICT	SOUTH BAČKA DISTRICT
Initial number of images	15,237 (100%)	19,580 (100%)	20,985 (100%)
Removed images with irrelevant tags	758 (4.97%)	1,832 (9.36%)	1,474 (7.02%)
Removed images with no tags (1st round)	7,466 (49%)	7,654 (39.09%)	9,814 (46.77%)
Total images removed (1st round)	8,224 (53.97%)	9,486 (48.45%)	11,288 (53.79%)
Total images remained (1st round)	7,013 (46.02%)	10,094 (51.55%)	9,697 (46.21%)
Total images removed (2nd round)	3,763 (24.69%) (53.66% of the 2nd round)	7,047 (35.99%) (69.81% of the 2nd round)	5,907 (28.15%) (60.92% of the 2nd round)
Total images remained (2nd round)	3,250 (21.33%) (46.34% of the 2nd round)	3,047 (15.56%) (30.19% of the 2nd round)	3,790 (18.06%) (39.08% of the 2nd round)

Table 6. Summary table of the main numbers of images for each of the three sites

3. 2. 2. Description of the Datasets

During this study, we used six datasets. 3 of them were made after the first round of cleaning and other 3 after the second round of cleaning. The first and the second group of datasets are related to Cambridgeshire, Coimbra district and South Bačka district. Also, one dataset is consisted of one training and validation dataset and one testing dataset. In the first round of cleaning datasets, testing sets were consisted of 500 images, but in the second only of 400 images. This is due to reduced datasets' sizes after the second round of cleaning. In the next

sections, the datasets obtained after the second round of data cleaning are going to be presented.

3. 2. 2. 1. Cambridgeshire datasets

The second round Cambridgeshire training dataset has 2,850 images, followed by 400 images in testing dataset.

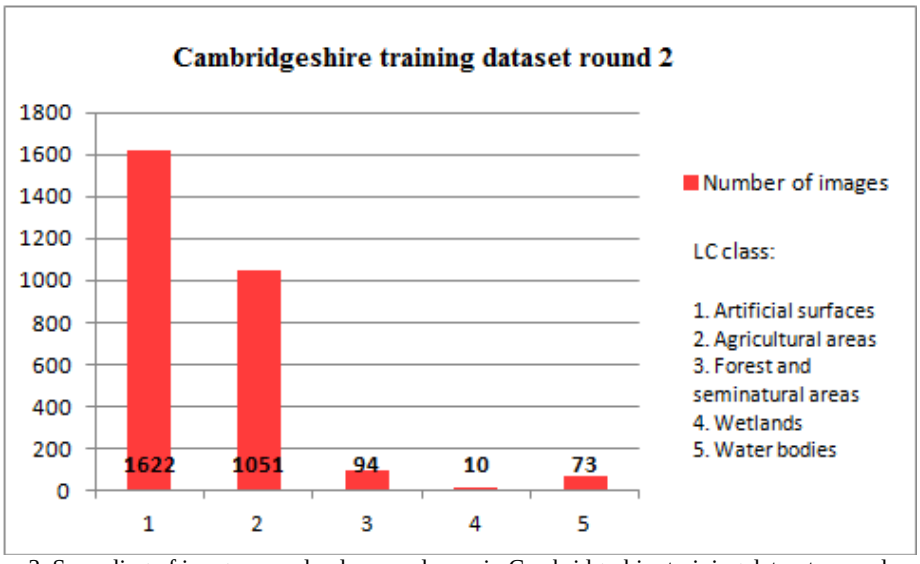

Figure 3. Spreading of images over land cover classes in Cambridgeshire training dataset second round

In Cambridgeshire training dataset images are spread like it follows (Figure 3):

1. Artificial surfaces - 1,622,

2. Agricultural areas - 1,051,

3. Forest and seminatural areas - 94,

4. Wetlands - 10,

5. Water bodies - 73.

In Cambridgeshire testing dataset images are spread like it follows (Figure 4):

1. Artificial surfaces - 237,

2. Agricultural areas - 150,

3. Forest and seminatural areas - 9,

4. Wetlands - 0,

5. Water bodies - 4.

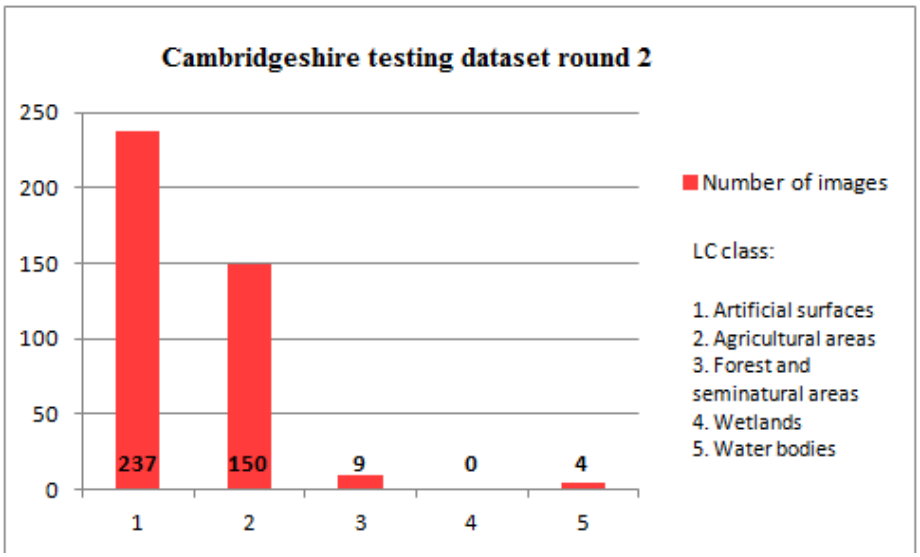

Figure 4. Spreading of images over land cover classes in Cambridgeshire testing dataset second round

In these datasets we can see extremely high numbers in first two classes comparing to three other classes. The biggest number of images is in class 1 (artificial surfaces). Little lower number is in class 2 (agricultural areas). Wetlands count really low number of images. In testing set there are 0 images in class 4 (wetlands). Class 3 (forest and seminatural areas) counts little bigger number, followed with class 5 (water bodies). Testing set was separated in completely random way, so, 0 images in class 4 only depict very low number of images in this class in the training dataset.

3. 2. 2. 2. Coimbra district datasets

The second round Coimbra district training dataset has 2,647 images, followed by 400 images in testing dataset.

In Coimbra district training dataset images are spread like it follows (Figure 5):

1. Artificial surfaces - 649,

2. Agricultural areas - 835,

3. Forest and seminatural areas - 1,029,

4. Wetlands - 18,

5. Water bodies - 116.

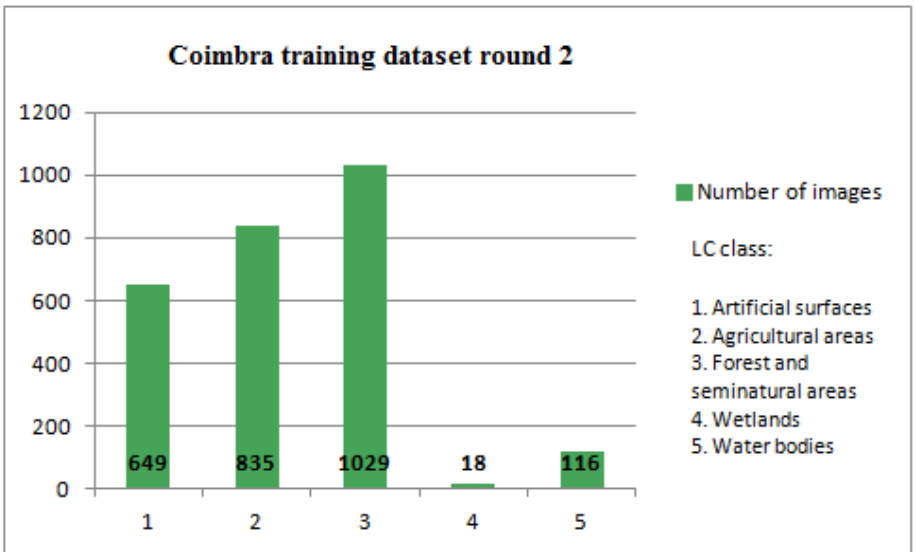

Figure 5. Spreading of images over land cover classes in Coimbra district training dataset second round

In Coimbra district testing dataset images are spread like it follows (Figure 6):

1. Artificial surfaces - 97,

2. Agricultural areas - 115,

3. Forest and seminatural areas - 172,

4. Wetlands - 2,

5. Water bodies - 14.

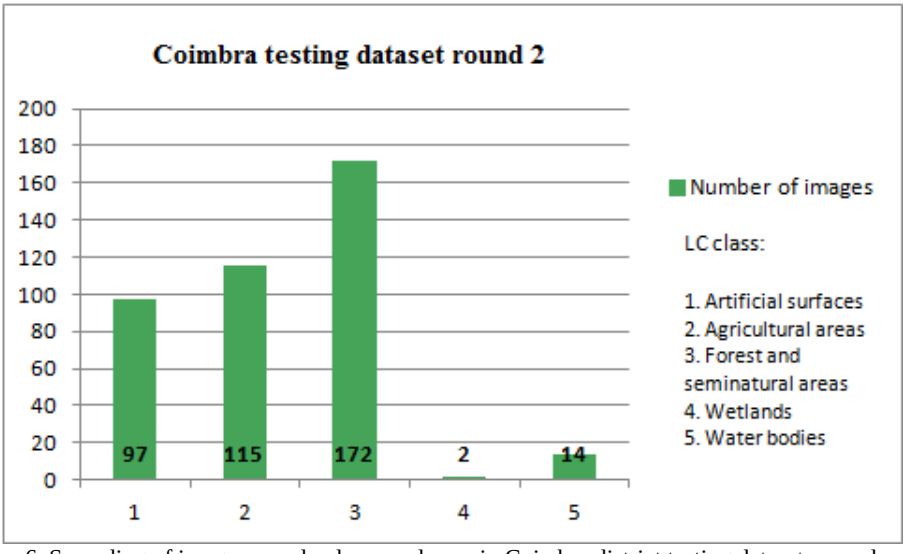

Figure 6. Spreading of images over land cover classes in Coimbra district testing dataset second round

Coimbra datasets are showing high number of images in class 3, little less in class 2 and at the end of these high number classes is class 1. The number of images which were taken on water bodies is quite small, while class 4 shows insignificant number of images. Later in the study we are going to see how models are confused with the first three classes. In other words, images with the same tags were taken on all of the three classes. This problem is significant and we couldn't see the solution for this.

3. 2. 2. 3. South Bačka district datasets

The second round South Bačka training dataset has 3,390 images, followed by 400 images in testing dataset.

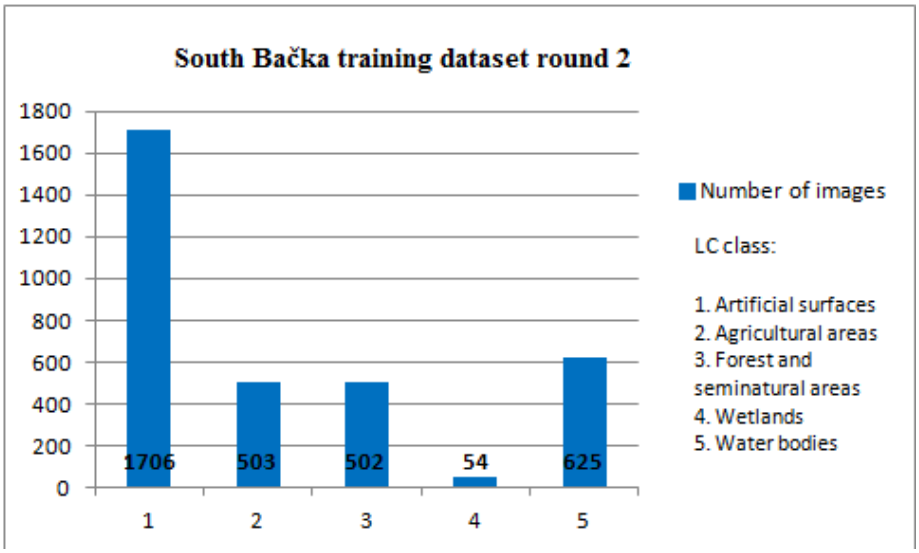

Figure 7. Spreading of images over land cover classes in South Bačka training dataset second round

In South Bačka training dataset images are spread like it follows (Figure 7):

1. Artificial surfaces - 1,706,

2. Agricultural areas - 503,

3. Forest and seminatural areas - 502,

4. Wetlands - 54,

5. Water bodies - 625.

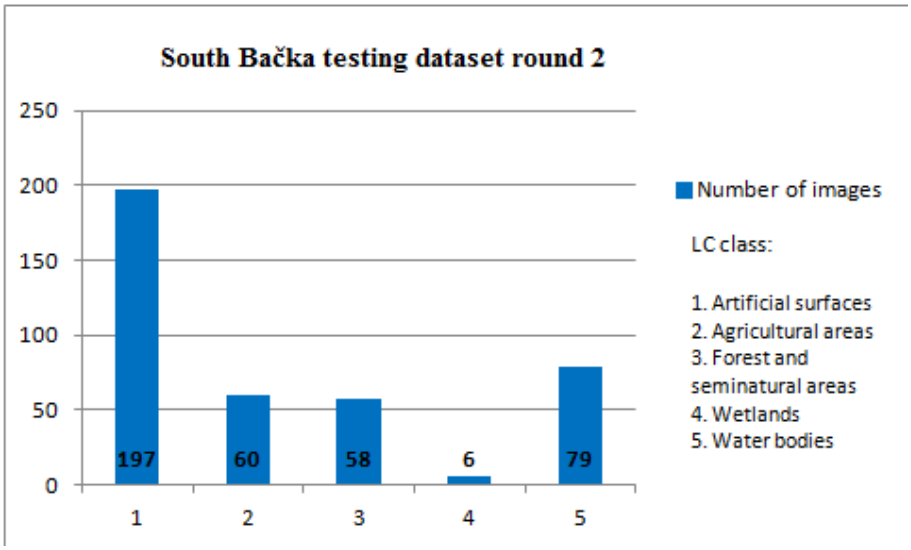

Figure 8. Spreading of images over land cover classes in South Bačka testing dataset second round

In South Bačka testing dataset images are spread like it follows (Figure 8):

1. Artificial surfaces - 197,

2. Agricultural areas - 60,

3. Forest and seminatural areas - 58,

4. Wetlands - 6,

5. Water bodies - 79.

South Bačka datasets are showing extremely high number of images which were taken in land cover class 1. In other words, people were mostly taking pictures inside the city. Surprisingly, the second class by number of images is class 5 - water bodies. This is not so surprising if we know that along the south border of the region is passing Danube River. The life on the river is very live, people spend a lot of time near the river and it is real touristic attraction. Also, Danube is wide enough that people are driving in boats and taking pictures from them or from the bridges. Those are the reasons why lot of images were geotagged exactly in class 5. Classes 2 and 3 are almost equal by their size. They are closely following class 5. Comparing to other two regions, wetlands are few times bigger, although they are still very small. This class will have almost no influences on the study.

3. 3. Data Analysis

In order to perform data analysis, we should get the data. The idea was to analyse tags from Panoramio. This was a good idea because the most of Panoramio images are orientated on

real geographical content. We downloaded the data using Panoramio public API (Application User Interface). The resulting datasets were manually cleaned afterwards.

These datasets were not extremely big, so we decided to clean it carefully in order to get more accurate results. Afterwards, it would be great if somebody develop automatic meanings of cleaning data which will be so effective like manual one.

As all downloaded images were geotagged, their coordinates and coordinates on CLC map allowed us to perform a "Spatial Join" in order to associate each location with their respective land cover class.

We had three different datasets, one for each of the next three regions: Cambridgeshire, Coimbra district and South Bačka district. We separated testing datasets, which were about 10% of the size of the initial dataset. Also, we performed basic analysis of the datasets, like is number of images per land cover class, comparing the same land cover classes in different datasets, etc.

The next step was to develop predictive models using the SAS software, specifically the SAS Enterprise Miner Workstation and the SAS Enterprise Guide. Because our target values were nominal, 5 methods were available to use in this study: Neural Networks, MBR (Memory Based Reasoning), Gradient Boosting, Regression and Decision Trees. After deep analysis of the significant amount of models developed, we concluded which method is the best and which are also not so bad. After building a model, we tested it on testing dataset. We also changed testing datasets to analyse if the results are similar.

Based on the results of predicted land cover classes on the testing datasets, we were building accuracy assessment tables. We based our analysis on concrete results on the real data. We analysed accuracy of every single land cover class and overall accuracy also. Based on this, we were able to conclude which method and model suits the best in the case of this study, which is not necessary indicator that this method will suit the best in all other cases and different kind of data.

3. 3. 1. SAS software

We decided to use SAS software in our study mainly because of its efficiency and our familiarity with it.

"Originally called "Statistical Analysis System," the SAS System is an integrated set of data management and decision support tools from the SAS Institute that runs on platforms from PCs to mainframes. It includes a complete programming language as well as modules for

spreadsheets, CBT, presentation graphics, project management, operations research, scheduling, linear programming, statistical quality control, econometric and time series analysis and mathematical, engineering and statistical applications. It also provides multidimensional data analysis (OLAP), query and reporting, EIS, data mining and data visualization (http://www.pcmag.com/encyclopedia/term/50809/sas-system, accessed: December, 2014)."

Specifically, in SAS software we used SAS Enterprise Miner Workstation, a part of SAS software, to build models and SAS Enterprise Guide to manage data. There are several methods inside SAS Enterprise Miner Workstation, but only 5 of them can be performed when the target value is nominal, which is our case. Those 5 methods are:

1. Neural Networks,
2. MBR (Memory Based Reasoning),
3. Gradient Boosting,
4. Regression,
5. Decision Trees.

"Artificial neural networks are computational methodologies that perform multifactorial analyses. Inspired by networks of biological neurons, artificial neural network models contain layers of simple computing nodes that operate as nonlinear summing devices. These nodes are richly interconnected by weighted connection lines, and the weights are adjusted when data are presented to the network during a "training" process. Successful training can result in artificial neural networks that perform tasks such as predicting an output value, classifying an object, approximating a function, recognizing a pattern in multifactorial data, and completing a known pattern. Many applications of artificial neural networks have been reported in the literature (Dayhoff and DeLeo, 2001)."

"Memory Based Reasoning (MBR) consists of variations on the nearest neighbor techniques. In its simplest formulation, MBR solves a new task by looking up examples of tasks similar to the new task and using similarity with these remembered solutions to determine the new solution (Masand et al., 1992)."

"Gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current "pseudo"-residuals by least squares at each iteration. The pseudo-residuals are the gradient of the loss functional being minimized, with

respect to the model values at each training data point evaluated at the current step (Friedman, 2002)."

"Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation (http://databases.about.com/od/datamining/g/regression.htm, accessed: December, 2014)."

"Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the values that are contained in that field (http://support.sas.com/publishing/pubcat/chaps/57587.pdf, accessed: December, 2014)."

3. 3. 2. Building a Predictive Model using SAS Software

SAS software was chosen because of its efficiency, our familiarity with it and nature of input variable, which is text. SAS Enterprise Miner has quite simple user-friendly interface, so, with minimal amount of coding, we could manage all the work until the end. Figures 9 and 10 are showing the structure of a predictive model in SAS Enterprise Miner. It is consisted of "nodes" and every single node and its function is going to be explained separately. One "node" represents one entity in SAS Enterprise Miner. It could be static, like dataset, or dynamic, like a node performing some action. Putting the nodes in appropriate and meaningful order and connecting them (running afterwards) is giving the certain result or output.

First node, "CAMBRIDGE" (Figure 9), is our initial training and validation dataset. The model is trained based on this dataset.

"StatExplore" and "MultiPlot" are optional nodes which create some statistics about the initial dataset. "MultiPlot" node enables exploring large volumes of data graphically.

The "Data Partition" node enables partitioning of datasets into training, test, and validation sets. This node uses simple random sampling, stratified random sampling, or a user-defined

partition to create training, test, or validation datasets (http://support.sas.com/documentation/cdl/en/emgsj/62040/HTML/default/viewer.htm#a003307717.htm, accessed: December, 2014).

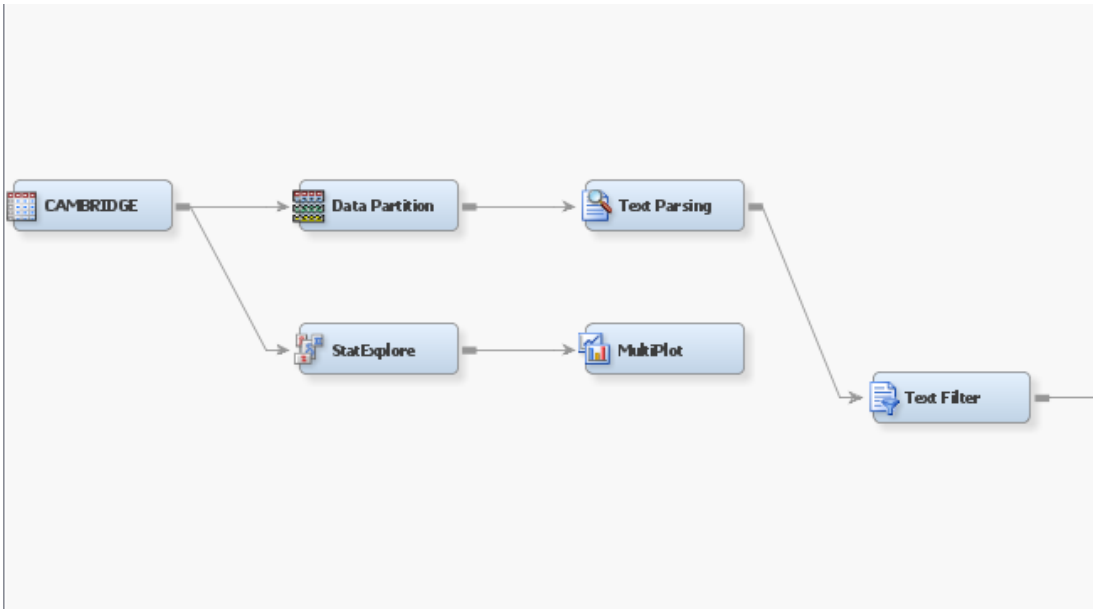

Figure 9. Structure of a predictive model in SAS Enterprise Miner (Part 1)

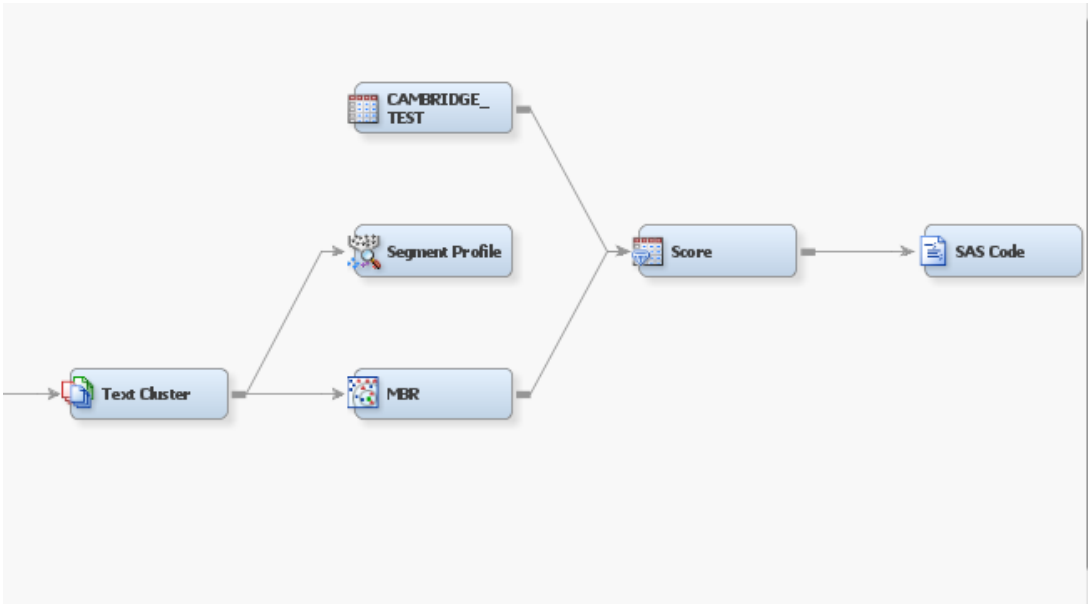

Figure 10. Structure of a predictive model in SAS Enterprise Miner (Part 2)

The "Text Parsing" node enables parsing a document collection in order to quantify information about the terms that are contained therein. It might be used with volumes of textual data such as e-mail messages, news articles, Web pages, research papers, and surveys (http://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs. pdf, accessed: December, 2014).

The "Text Filter" node can be used to reduce the total number of parsed terms or documents that are analysed. Therefore, it is possible to eliminate extraneous information so that only the most valuable and relevant information is considered (http://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf, accessed: December, 2014).

The "Text Cluster" node, presented in Figure 10, clusters documents into disjointed sets of documents and reports on the descriptive terms for those clusters. Two algorithms are available: 1) the Expectation Maximization algorithm that clusters documents with a flat representation and 2) the Hierarchical clustering algorithm that groups clusters into a tree hierarchy. Both approaches rely on the Singular Value Decomposition (SVD) to transform the original weighted, term-document frequency matrix into a dense but low dimensional representation (http://support.sas.com/ documentation/onlinedoc/txtminer/12.1/tmgs.pdf, accessed: December, 2014).

The "Segment Profile" node is used to provide a better idea of what makes each segment unique or at least different from the population. The node generates various reports that aid in exploring and comparing the distribution of these factors within the segments and population (http://support.sas.com/documentation/cdl/en/tmgs/ 62416/HTML/default/viewer.htm#n1gpxb3f6uwsu8n149yftcyonlql.htm, accessed: December, 2014). This node is not mandatory.

The "MBR" node, shown in Figure 10, is a node where the data mining method is chosen among the 5 available options as stated before in this section: Regression, Neural Networks, MBR, Gradient Boosting and Decision Trees.

The "CAMBRIDGE_TEST" node is the node where the target value should be predicted. This is a testing dataset node, just we didn't partition initial dataset into training, validation and testing datasets, but we used testing dataset separately like scoring node. This was performed in order to be able to see the "scores", the results of a prediction.

The "Score" node is a node that assign predicted values to target variable in the scoring dataset based on predictive model which has been built in previous steps.

The "SAS Code" node is used to apply different codes into the workflow. It could be any codes supported by the SAS software. In our case, we have used "SAS Code" node to export the results into a format more suitable for visualisation and further analysis.

After building the model, an accuracy assessment was performed to evaluate the model through the development of confusion matrices. Building confusion matrices were done by using a code which were comparing predicted value with real value and assigning a number from 1 to 25 to the certain feature. According to this number, a certain feature is assigned to the one of the cells from the 5x5 table. This table represents basis of the confusion matrices (5 real and 5 predicted land cover classes).

3. 3. 3. Accuracy Assessment

Accuracy assessment is a general term for comparing a new dataset with a reference dataset assumed to be true, in order to determine the accuracy of the classification process. Evaluation of the accuracy can be done using an error matrix sometimes called confusion matrix (Senseman et al., 1995; Foody, 2002).

Error Matrices or confusion matrices are composed by a square array of numbers laid out in rows and columns which expresses the number of sample units assigned to a particular category relative to the actual category for which we have true values. The columns normally represent the reference data, while the rows indicate the data to be assessed. Most of the classification accuracy measurements are derived from an error matrix. However, the most popular one is the correctly classified cases in a percentage (Senseman et al., 1995; Maingi et al., 2002).

In our study, we used confusion matrices to perform the accuracy assessment. As far as we had true values in testing dataset, we could compare them with predicted values in the same dataset, but different column. Based on this, we were able to create confusion matrices as described in previous section. After creating confusion matrices, we ensured all necessary conditions for performing analysis of the results.

In order to make our results more clear, it is needed to provide some explanations about confusion matrices (Figure 11) that we have used in our analysis.

In the first row it is written which algorithm is used. The classes from 1 to 5 in the first column are true classes. The classes from 1 to 5 in the second row are predicted classes.

Cross line bold numbers (CC) in confusion matrices are correctly classified images.

Value N represents total number of images in each class.

"Class Acc." column represents percentage of correctly classified images in a single class. "Av." value at the bottom is average class accuracy.

CONFUSION MATRIX (NEURAL NETWORKS)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	CC						
2		CC					
3			CC				
4				CC			
5					CC		
Total (X)							Av.
CC/X						Av. CC/X	
P=X/N						Σ|P-100|/C	T. Im.
D=X-N						Σ|D|	Acc.

Figure 11. Example of a confusion matrix

Value X represents all images classified in one class, independently if they are correct or no.

"CC/X" is ratio between correctly classified features and value X. "Av. CC/X" is average value of previous ones.

Value P represents value X divided with value N. We can depict percentage of over or under classifying of certain class with this value. "Σ|P-100|/C" at the end of previous row represents averagely how much percentage is over or under classified in any of the classes.

Value D represents difference of the classified images in a certain class with the total number of images in this class. This value can be positive or negative. It depicts how many images are over or under classified like certain class. At the end, we can sum everything like absolute numbers (Σ|D|) to see how many images are classified over or under true number of images in every class in total.

Values P and D are used to depict an algorithm's ability to distinguish different classes. The most of the algorithms are showing tendency to classify the most of the images as the most frequent class, neglecting less frequent classes. Especially because of the occurrence of this phenomenon, these values were established.

"T. Im." represents total number of images tested. "Acc." is overall accuracy.

With using the values described above, knowing datasets and overall workflow, we performed analysis, discussed and explained our results and answered our research questions.

CHAPTER FOUR

4. RESULTS AND DISCUSSION

This chapter will discuss the results obtained by applying the methodology defined in chapter Data and Methodology. The results are divided in two parts according to the data cleaning process. The first part corresponds to the first round of data cleaning, and the second part to the second round of data cleaning.

4. 1. The First Round of the Data Cleaning Results

We performed some rough data cleaning and we got some results. These results will be presented in this section. In building predictive models and analysing them, we used five methods, performed on three different areas. So, this is fifteen different combinations of methods and areas. In other words, this is fifteen confusion matrices presented in this section. We are going to explain our first results shortly.

4. 1. 1. Cambridgeshire dataset results - the first round

Cambridgeshire testing set has 500 images. The majority of them (303) are in class 1. Class 4 doesn't contain images at all. This is because the number of images in wetlands is very low. Classes 3 and 5 are almost insignificant, but distinguishing them could show sensitivity of a certain method. Class 2 contains 183 images.

CONFUSION MATRIX (NEURAL NETWORKS)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**250**	53	0	0	0	303	82%
2	123	**60**	0	0	0	183	33%
3	7	2	**0**	0	0	9	0%
4	0	0	0	**0**	0	0	N/A
5	4	1	0	0	**0**	5	0%
Total (X)	384	116	0	0	0	**310**	Av. 28.8%
CC/X	65%	51%	N/A	N/A	N/A	Av. CC/X = 58%	
P=X/N	126%	63%	0%	N/A	0%	Σ|P-100|/C=66%	T. Im.:**500**
D=X-N	81	-67	-9	0	-5	Σ|D| = 162	**Acc. 62%**

Table 7. Neural Networks model on Cambridgeshire dataset (1st round)

5 methods were applied on this dataset in order to build predictive model. The results of the scoring (predicting) of testing dataset are as follows:

1. Neural Networks method achieved 62% of overall accuracy (Table 7). Accuracy of class 1 is quite high (82%), while the second frequent class (class2) has accuracy of 33%. Classes 3 and 5 have accuracy of 0%, which is not quite surprising because very small amount of images were in these datasets. We can conclude that 62% of overall accuracy was the consequence of the high accuracy of the most frequent class (class1).

CONFUSION MATRIX (GRADIENT BOOSTING)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**269**	34	0	0	0	303	89%
2	130	**53**	0	0	0	183	29%
3	7	2	**0**	0	0	9	0%
4	0	0	0	**0**	0	0	N/A
5	5	0	0	0	**0**	5	0%
Total (X)	411	89	0	0	0	**322**	Av. 29.5%
Corr.Class./X	65%	60%	N/A	N/A	N/A	Av. CC/X = 62.5%	
P=X/N	137%	49%	0%	N/A	0%	Σ|P-100|/C=72%	T. Im.:**500**
D=X-N	108	-94	-9	0	-5	Σ|D| = 216	**Acc. 64.4%**

Table 8. Gradient Boosting model on Cambridgeshire dataset (1st round)

CONFUSION MATRIX (MBR)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**248**	50	5	0	0	303	82%
2	108	**76**	0	0	0	183	42%
3	3	5	**1**	0	0	9	11%
4	0	0	0	**0**	0	0	N/A
5	3	2	0	0	**0**	5	0%
Total (X)	362	133	6	0	0	**325**	Av. 43.8%
Corr.Class./X	68%	57%	17%	N/A	N/A	Av. CC/X = 47%	
P=X/N	119%	44%	67%	N/A	0%	Σ|P-100|/C=52%	T. Im.:**500**
D=X-N	59	-50	-3	0	-5	Σ|D| = 117	**Acc. 65%**

Table 9. MBR model on Cambridgeshire dataset (1st round)

2. Gradient Boosting method achieved 64.4% of overall accuracy (Table 8). Accuracy of class 1 is 89%. Class 2 has 29% of accuracy. Classes 3 and 4 have 0%. Although overall accuracy is higher than in previous case, class 2 accuracy is lower. That means that the model determined better the most frequent class, but it is not so efficient with other classes.

3. MBR method achieved 65% of overall accuracy (Table 9). Accuracy of class 1 is 82%, but class 2 has 42% of accuracy. Even in class 3 there is one image correctly classified. Accuracy of this class is 11%. This model didn't show the best accuracy result of class 1, but it showed much higher accuracy in other classes. This is very important to notice. We will see later that some models have almost 100% of accuracy in the most frequent class, but very low accuracy of other classes. This gave them very high overall accuracy, but in reality, they are bad models because they couldn't determine any other class, just classifying all images in the most frequent class.

4. Regression method achieved 60.4% of overall accuracy (Table 10). Accuracy of class 1 is quite high (90%), but accuracy of class 2 is quite low (16%). This model managed to correctly classify 1 image in class 5, so, accuracy of this class is 20%. This model has tendency to classify majority of features like class 1. That is the reason why it achieved quite high overall accuracy.

CONFUSION MATRIX (REGRESSION)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**272**	30	0	0	1	303	90%
2	154	**29**	0	0	0	183	16%
3	7	0	**0**	0	2	9	0%
4	0	0	0	**0**	0	0	N/A
5	4	0	0	0	**1**	5	20%
Total (X)	437	59	0	0	4	**302**	Av. 31.5%
Corr.Class./X	62%	49%	N/A	N/A	25%	Av. CC/X = 45%	
P=X/N	144%	32%	0%	N/A	80%	Σ|P-100|/C=58%	T. Im.:**500**
D=X-N	134	-124	-9	0	-1	Σ|D| = 268	**Acc. 60.4%**

Table 10. Regression model on Cambridgeshire dataset (1st round)

5. Decision Trees achieved 60.4% of overall accuracy (Table 11). Accuracy of class 1 is extremely high - 98%, but accuracy of class 2 is only 3%. This model managed to correctly classify 3 images in class 5, so, accuracy of this class is 60%. This model is the best example of a model which has quite high overall accuracy like consequence of classifying almost all images like the most frequent class. Even if we get the best overall accuracy with Decision Trees method, this is still very bad model in this case.

These are results from Cambridgeshire dataset. We are going to present the results of other two datasets and see if there is some similarity between the same methods.

Overall accuracy is not high enough to prove our research question, so, our decision was to perform more detail cleaning of data. We are going to present these results later.

CONFUSION MATRIX (DECISION TREES)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**297**	5	0	0	1	303	98%
2	181	**5**	0	0	0	186	3%
3	7	0	**0**	0	2	9	0%
4	0	0	0	**0**	0	0	0%
5	1	1	0	0	**3**	5	60%
Total (X)	486	11	0	0	6	**302**	Av. 32.2%
Corr.Class./X	61%	45%	N/A	N/A	50%	Av. CC/X = 52%	
P=X/N	160%	6%	0%	N/A	120%	Σ|P-100|/C=68%	T. Im.:**500**
D=X-N	183	-175	-9	0	1	Σ|D| = 368	**Acc. 60.4%**

Table 11. Decision Trees model on Cambridgeshire dataset (1st round)

4. 1. 2. Coimbra district dataset results - the first round

Coimbra district testing set has 500 images. The majority of them (208) are in class 3, followed by class 2 (158) and class 1 (108). There are only 18 images in class 5 and 8 images in class 4. Number of images in class 5 is not so high, despite the presence of the ocean. This is probably because people were taking pictures of the ocean from the land.

The results of applied 5 methods on this dataset are as follows:

CONFUSION MATRIX (NEURAL NETWORKS)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**40**	2	66	0	0	108	37%
2	23	**19**	116	0	0	158	12%
3	24	16	**168**	0	0	208	81%
4	2	1	5	**0**	0	8	0%
5	2	1	15	0	**0**	18	0%
Total (X)	91	39	370	0	0	**227**	Av. 26%
Corr.Class./X	44%	49%	45%	N/A	N/A	Av. CC/X = 46%	
P=X/N	84%	25%	178%	0%	0%	Σ|P-100|/C=74%	T. Im.:**500**
D=X-N	-17	-119	162	-8	-18	Σ|D| = 324	**Acc. 45.4%**

Table 12. Neural Networks model on Coimbra district dataset (1st round)

1. Neural Networks method achieved 45.4% of overall accuracy (Table 12). Class 3, as the most frequent class, has accuracy of 81%, while class 2 has 12%. This is probably because of similarity between these two classes, so, many of images from class 2 are classified like class 3. Class 1 has 37% of accuracy, while classes 4 and 5 don't contain any image classified so.

2. Gradient Boosting method achieved 42.8% of overall accuracy (Table 13). Class 3 has 84% of accuracy, while classes 1 and 2 have 14% of accuracy, both of them. In this case this method classified not only many of class 2 images like class 3, but also many of class 1 images. Classes 4 and 5 don't have any image classified in their classes.

CONFUSION MATRIX (GRADIENT BOOSTING)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**15**	9	84	0	0	108	14%
2	8	**22**	127	0	0	158	14%
3	11	25	**172**	0	0	208	83%
4	2	2	4	**0**	0	8	0%
5	0	2	16	0	**0**	18	0%
Total (X)	36	60	403	0	0	**214**	Av. 22.2%
Corr.Class./X	42%	37%	43%	N/A	N/A	Av. CC/X = 41%	
P=X/N	33%	38%	194%	0%	0%	Σ|P-100|/C=85%	T. Im.:**500**
D=X-N	-72	-98	195	-8	-18	Σ|D| = 391	**Acc. 42.8%**

Table 13. Gradient Boosting model on Coimbra district dataset (1st round)

3. MBR method achieved 40% of overall accuracy (Table 14). Class 3 has 47% of accuracy, but classes 1 and 2 are, with comparing to previous methods, quite similarly highly accurate, 42% and 35%, respectively. Despite low overall accuracy, this is one more proof that MBR method is able to distinguish classes well and probable potential of usability of this method in our study. Classes 4 and 5 have 0% of accuracy.

CONFUSION MATRIX (MBR)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**46**	24	38	0	1	108	42%
2	29	**56**	69	0	4	158	35%
3	58	48	**98**	0	5	208	47%
4	2	3	3	**0**	0	8	0%
5	2	6	10	0	**0**	18	0%
Total (X)	137	137	218	0	10	**200**	Av. 24.8%
Corr.Class./X	34%	41%	45%	N/A	0%	Av. CC/X = 30%	
P=X/N	127%	87%	105%	0%	56%	Σ|P-100|/C=38%	T. Im.:**500**
D=X-N	29	-21	10	-8	-8	Σ|D| = 76	**Acc. 40%**

Table 14. MBR model on Coimbra district dataset (1st round)

CONFUSION MATRIX (REGRESSION)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**24**	6	78	0	0	108	22%
2	18	**14**	126	0	0	158	9%
3	20	22	**166**	0	0	208	80%
4	3	2	3	**0**	0	8	0%
5	1	2	15	0	**0**	18	0%
Total (X)	66	46	388	0	0	**204**	Av. 22.2%
Corr.Class./X	36%	30%	43%	N/A	N/A	Av. CC/X = 36%	
P=X/N	61%	29%	186%	0%	0%	Σ|P-100|/C=79%	T. Im.:**500**
D=X-N	-42	-112	180	-8	-18	Σ|D| = 360	**Acc. 40.8%**

Table 15. Regression model on Coimbra district dataset (1st round)

4. Regression method achieved 40.8% of overall accuracy (Table 15). Class 3 has 80% of accuracy, while class 2 has extremely low accuracy, only 9%. This is because many of images from class 2 were classified like class 3. Class 1 has 22%, with the same phenomenon. Classes 4 and 5 have accuracy of 0%.

CONFUSION MATRIX (DECISION TREES)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**8**	0	100	0	0	108	7%
2	3	**3**	152	0	0	158	2%
3	2	3	**203**	0	0	208	98%
4	0	0	8	**0**	0	8	0%
5	0	0	18	0	**0**	18	0%
Total (X)	13	6	481	0	0	**214**	Av. 21.6%
Corr.Class./X	62%	50%	42%	N/A	N/A	Av. CC/X = 51%	
P=X/N	12%	4%	231%	0%	0%	Σ|P-100|/C=103%	T. Im.:**500**
D=X-N	-95	-152	273	-8	-18	Σ|D| = 546	**Acc. 42.8%**

Table 16. Decision Trees model on Coimbra district dataset (1st round)

5. Decision Trees method achieved 42.8% of overall accuracy (Table 16). Class 3 has extremely high accuracy (98%), while classes 1 and 2 extremely low (7% and 2%). This shows how Decision Trees method has great tendency to classify almost all features in the most frequent class. Its overall accuracy is only illusion, because almost all of accurately

classified images belong to the class 3, which is the most frequent class. Classes 4 and 5 have 0% of accuracy.

The results of Coimbra dataset models were not satisfied at all. One part of this is unfamiliarity with Portuguese language (which was important during data cleaning) and other part is really messy data, like, for example, several images with completely the same tags belong to three different classes, etc.

4. 1. 3. South Bačka district dataset results - the first round

South Bačka district testing set has 500 images. The majority of them (283) are in class 1. Class 2 contains 107 images. Number of images in class 5 is surprisingly high (59), followed by 49 images in class 3. Number of images in class 5 is high because of presence of Danube River and people's attitude to spend their time on it. Class 4 contains only 2 images.

The results of 5 predictive models, gotten by 5 different methods, are as follows:
1. Neural Networks method achieved 62.4% of overall accuracy (Table 17). Class 1, as the most frequent class, has 90% of accuracy. Classes 3 and 5 have 33% and 34% of accuracy, respectively. Class 2 has only 18% of accuracy. The areas around settlements are predominantly agricultural. This might be reason for so high error in classifying of class 2 - it is mostly classified like class 1. Surprisingly, class 4 has both images correctly classified, so, accuracy of this class is 100%.

CONFUSION MATRIX (NEURAL NETWORKS)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**255**	7	6	0	15	283	90%
2	78	**19**	3	0	7	107	18%
3	25	5	**16**	0	3	49	33%
4	0	0	0	**2**	0	2	100%
5	35	4	0	0	**20**	59	34%
Total (X)	393	35	25	2	45	**312**	Av. 55%
Corr.Class./X	65%	54%	64%	100%	44%	Av. CC/X = 65%	
P=X/N	139%	33%	51%	100%	76%	Σ|P-100|/C=36%	T. Im.:**500**
D=X-N	110	-72	-24	0	-14	Σ|D| = 220	**Acc. 62.4%**

Table 17. Neural Networks model on South Bačka district dataset (1st round)

CONFUSION MATRIX (GRADIENT BOOSTING)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**266**	4	7	0	6	283	94%
2	76	**25**	4	0	2	107	23%
3	25	3	**19**	0	2	49	39%
4	2	0	0	**0**	0	2	0%
5	41	3	0	0	**15**	59	25%
Total (X)	410	35	30	0	25	**325**	Av. 36.1%
Corr.Class./X	65%	71%	63%	N/A	60%	Av. CC/X = 65%	
P=X/N	145%	33%	61%	0%	42%	Σ|P-100|/C=62%	T. Im.:**500**
D=X-N	127	-72	-19	-2	-34	Σ|D| = 254	**Acc. 65%**

Table 18. Gradient Boosting model on South Bačka district dataset (1st round)

2. Gradient Boosting method achieved 65% of overall accuracy (Table 18). Class 1 has very high level of accuracy (94%). Class 3 has 39%, class 5 - 25% and class 2 - 23%. This method showed very good characteristics in this case. Only class 5 is worse classified than in previous method. However, the images in class 4 are not correctly classified also.

CONFUSION MATRIX (MBR)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**210**	25	5	0	43	283	74%
2	68	**29**	3	0	7	107	27%
3	15	13	**17**	0	4	49	35%
4	0	0	0	**2**	0	2	100%
5	30	4	0	1	**24**	59	41%
Total (X)	323	71	25	3	78	**282**	Av. 55.4%
Corr.Class./X	65%	41%	68%	67%	31%	Av. CC/X = 54%	
P=X/N	114%	66%	51%	150%	132%	Σ|P-100|/C=36%	T. Im.:**500**
D=X-N	40	-36	-24	1	19	Σ|D| = 120	**Acc. 56.4%**

Table 19. MBR model on South Bačka district dataset (1st round)

3. MBR method achieved 56.4% of overall accuracy (Table 19). Class 1 has 74% of accuracy, which is really not quite high. Class 5 has 41% of accuracy. Classes 2 and 3 have 27% and 35% of accuracy, respectively. This method classified very well images in class 4, 100% of accuracy. This method also showed higher accuracy in class 5 than previous

methods. Once again, we confirmed that this method is very good in distinguishing classes, although, not so good in overall accuracy in this case.

CONFUSION MATRIX (REGRESSION)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**266**	7	5	0	5	283	94%
2	82	**17**	5	0	3	107	16%
3	29	0	**17**	0	3	49	35%
4	0	0	0	**2**	0	2	100%
5	43	3	1	0	**12**	59	20%
Total (X)	420	27	28	2	23	**315**	Av. 53%
Corr.Class./X	63%	63%	61%	100%	52%	Av. CC/X = 68%	
P=X/N	148%	25%	57%	100%	39%	Σ|P-100|/C=45%	T. Im.:**500**
D=X-N	137	-80	-21	0	-36	Σ|D| = 274	**Acc. 63%**

Table 20. Regression model on South Bačka district dataset (1st round)

4. Regression method achieved 63% of overall accuracy (Table 20). Class 1 has very high accuracy (94%). But, this method is showing tendency to classify majority of images in class 1, like the most frequent class. So, accuracy of other classes are much lower (class 2 - 16%, class 3 - 35%, class 5 - 20%). Contradictory, it classified both images from class 4 correctly.

CONFUSION MATRIX (DECISION TREES)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**277**	3	3	0	0	283	98%
2	83	**15**	5	0	0	103	14%
3	30	1	**18**	0	0	49	37%
4	0	0	2	**0**	0	2	0%
5	57	2	0	0	**0**	59	0%
Total (X)	447	21	28	0	0	**313**	Av. 29.8%
Corr.Class./X	62%	71%	64%	N/A	N/A	Av. CC/X = 66%	
P=X/N	158%	20%	57%	0%	0%	Σ|P-100|/C=76%	T. Im.:**500**
D=X-N	164	-82	-21	-2	-59	Σ|D| = 328	**Acc. 62.6%**

Table 21. Decision Trees model on South Bačka district dataset (1st round)

5. Decision Trees method achieved 62.6% of overall accuracy (Table 21). Class 1 has extremely high level of accuracy (98%) and this is main reason for quite satisfying overall accuracy (if we compare with other methods). But, other classes have quite low level of accuracy. Classes 4 and 5 have 0% level of accuracy each. Class 2 has 14%. Class 3 has

37%. Tendency for classifying huge majority of images in the most frequent class is evident, and it is proven at the case with this method.

Even in this moment, we can say that Decision Trees method is not good for this type of data. Even Regression method is showing some tendencies similar like Decision Trees, and this is classifying the majority of images in the most frequent class and consequently getting good overall accuracy.

The results of predictive models made on this dataset are better, but still not satisfying enough. Because of this, we performed the second round of data cleaning. These results are going to be presented in next section.

4. 2. The Second Round of Data Cleaning Results

The second round of data cleaning gave better results. Testing the models showed that data cleaning is one of the most important steps, maybe the most important one. In the next three sections, we presented the best predictive models that we could build for our three different areas with three different speaking languages.

We decided not to present Regression and Decision Trees models because they have tendency to classify the majority of instances in the most frequent class. Although they have good overall accuracy, they are not good methods for this kind of data.

Explaining one by one model with the best performances built on three datasets is going to be the way of presenting the results. Later on, comparing them and discuss their performances will show which method and model gave the best results.

4. 2. 1. Cambridgeshire dataset results - the second round

Cambridgeshire testing set has 400 images. The most of them are in class 1 (237), followed by class 2 (150). Class 3 and 5 contain 9 and 4 images, respectively, while class 4 has no images.

Our model with the best results which we got using Neural Networks method (Table 22) has 65.75% of the overall accuracy.

Class 1 has 84% of accuracy. In other words, 198 images were classified correctly out of 237 in total. Class 2 has little lower accuracy (42%), while class 3 has only 22% of accuracy. But class 3 contains only 9 images and 2 of them were classified correctly. Class 4 contains no images, so we cannot say anything about it. Class 5 contains 4 images and no one of them was classified correctly, so we have accuracy of 0% in this class.

We can see that a high number of images from class 2 were classified as class 1. Also, not so small number of the images from class 1 was classified as class 2. These two classes are similar and the model had difficulties to distinguish them, probably because the class 2 (agricultural areas) is usually surrounding class 1 (artificial surfaces). The reason could be that most of such images were taken by standing on one class and picturing and tagging the other (Zielstra, 2013).

CONFUSION MATRIX (NEURAL NETWORKS)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**198**	39	0	0	0	237	84%
2	83	**63**	4	0	0	150	42%
3	7	0	**2**	0	0	9	22%
4	0	0	0	**0**	0	0	N/A
5	4	0	0	0	**0**	4	0%
Total (X)	292	102	6	0	0	**263**	Av. 37%
CC/X	68%	62%	33%	N/A	N/A	Av. CC/X = 54%	
P=X/N	123%	68%	67%	N/A	0%	Σ|P-100|/C=47%	T. Im.:**400**
D=X-N	55	-48	-3	0	-4	Σ|D| = 110	**Acc. 65.75%**

Table 22. Confusion matrix of Neural Networks predictive model built on Cambridgeshire dataset

Value P is higher in class 1, while in all other classes is lower than 100%. This means that more images were classified in class 1 then it really exists in this class. In other classes, less number of images was classified like these classes than it really exists in these classes. All this is depicted with value D also, but in absolute numbers.

The model with the best results made with Gradient Boosting method (Table 23) has 65.25% of the overall accuracy.

Class 1 showed higher accuracy than the previous model, which is 87% or, in other words, 206 images were correctly classified out of 237 in total. But other classes showed lower accuracy. Hence, class 2 had 37% of accuracy, while class 3 had 0%. There were no classified images at all in classes 3, 4 and 5.

This behaviour is similar to the behaviour of the Decision Trees method and until some extent Regression, which is not so surprising once the Gradient Boosting algorithm is based on Decision Trees (Friedman, 2001).

Like in previous case, value P is higher than 100% in class 1, like the most frequent class, and lower in class 2, while classes 3 and 5 don't have any images classified like these

classes. In absolute numbers, there were 62 images classified as class 1 more than class 1 really contains. In class 2, 60 images were classified less as class 2 than this class really contains.

CONFUSION MATRIX (GRADIENT BOOSTING)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**206**	30	0	0	1	237	87%
2	85	**55**	0	0	0	150	37%
3	6	3	**0**	0	0	9	0%
4	0	0	0	**0**	0	0	N/A
5	2	2	0	0	**0**	4	0%
Total (X)	299	90	0	0	0	**261**	Av. 31%
CC/X	69%	61%	N/A	N/A	N/A	Av. CC/X = 65%	
P=X/N	126%	60%	0%	N/A	0%	Σ|P-100|/C=66%	T. Im.:**400**
D=X-N	62	-60	-9	0	-4	Σ|D| = 135	**Acc. 65.25%**

Table 23. Confusion matrix of Gradient Boosting predictive model built on Cambridgeshire dataset

The model with the best results made with MBR method (Table 24) has 71.75% of the overall accuracy.

In the case of MBR method made model, class 1 didn't show the highest accuracy if we compare with previous two models (83%), but other classes are much more accurate. Class 2 has 57% of accuracy. This is much higher than the value obtained using the previous two methods, which was 42% for Neural Networks model and 37% for Gradient Boosting model. Class 3 is also more accurate, 33%, while class 5 has 50% of accuracy. In the previous two models, class 5 had 0% of accuracy.

This method much better distinguishes classes. Although accuracy in the most frequent class is not the highest among these three models, other classes are much better classified. So, overall accuracy is the highest in this case. Even if overall accuracy is not the highest, it can be said that this method has the best abilities for classifying not only the most frequent class, but also the small ones.

Value P performed better in the case of the MBR based model. Only 11% more classified as class 1 than it really exists, 16% less in class 2 and 22% less in class 3 (classified in these classes than it really exists). Class 5 contains 4 classified images (two correctly and two incorrectly) which is equal with the number of images in this class, so it means the value P is 100%.

CONFUSION MATRIX (MBR)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**196**	38	1	0	2	237	83%
2	61	**86**	3	0	0	150	57%
3	4	2	**3**	0	0	9	33%
4	0	0	0	**0**	0	0	N/A
5	2	0	0	0	**2**	4	50%
Total (X)	263	126	7	0	4	**287**	Av. 56%
CC/X	74%	68%	43%	N/A	50%	Av. CC/X = 59%	
P=X/N	111%	84%	78%	N/A	100%	Σ|P-100|/C=12%	T. Im.:**400**
D=X-N	26	-24	-2	0	0	Σ|D| = 52	**Acc. 71.75%**

Table 24. Confusion matrix of MBR predictive model built on Cambridgeshire dataset

Among these three methods, MBR method expressed itself like the best method in building predictive models in text mining. We would give an advantage to this method among the others even if the overall accuracy is not the highest, because of its ability to distinguish classes better than the other methods. But, also the overall accuracy is the highest in doing prediction with this model. So, estimation of this model is assured.

4. 2. 2. Coimbra district dataset results - the second round

Coimbra district testing set has 400 images. In this case, class 3 is the most frequent class with 162 images. Second and first class are following class 3 with 115 and 97 images, respectively. Class 5 has 14 images, while class 4 contains only 2 images.

Neural Networks method gave the model (Table 25) with the best result of 46% of the overall accuracy. This is the best model which we could get with this method.
The best accuracy among classes showed the class 3, which was expected if we know this is the most frequent class. 114 images out of 162 were classified correctly. The class 2 has 34% of accuracy, while the class 1 has 32%. Classes 4 and 5 have 0% of accuracy.

We can see big number of images from classes 1 and 2 classified like the class 3. Also, not small number of images from the class 3 was classified like classes 1 and 2. This could be explained with land cover type which predominates in Coimbra district. Forests, like major land cover type, are surrounding towns and villages, and they are mixed with agricultural areas. So, we have similar case like in Cambridgeshire, which is taking photos of a sight from one land cover class from standing point on another land cover class (Zielstra, 2013).

Classes 4 and 5 have no classified images at all. The low level of the accuracy of the class 5 could be explained with people's habit to take photos of the ocean from the land (Estima, Fonte and Painho, 2014) i.e. from the other land cover classes.

Low overall accuracy shows how important is data cleaning. Unfamiliarity with Portuguese language caused this low accuracy. This was a good experiment for checking how important is cleaning of data in the process of the building of a predictive model in the text mining.

Value P is very high in the class 3, like the most frequent class. 46% of images were classified as the class 3 more than it really exists in this class. Classes 1 and 2 have 34% and 22% (respectively) less classified images like these classes than there really exists. Classes 4 and 5 has 100% less classified images as these classes, which means there are no images classified like these classes. In absolute numbers, it is worth mentioning that in the class 3 we have 74 images classified more that this class really contains.

CONFUSION MATRIX (NEURAL NETWORKS)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**31**	22	44	0	0	97	32%
2	9	**39**	67	0	0	115	34%
3	21	27	**114**	0	0	162	70%
4	1	1	0	**0**	0	2	0%
5	2	1	11	0	**0**	14	0%
Total (X)	64	90	236	0	0	**184**	Av. 27%
CC/X	48%	43%	48%	N/A	N/A	Av. CC/X = 46%	
P=X/N	66%	78%	146%	0%	0%	Σ|P-100|/C=60%	T. Im.:**400**
D=X-N	-33	-25	74	-2	-14	Σ|D| = 148	**Acc. 46%**

Table 25. Confusion matrix of Neural Networks predictive model built on Coimbra district dataset

Gradient boosting method gave a model with the best result (Table 26) which is 47.25% of the overall accuracy. It showed even better performances than Neural Networks method in this case. This was not the case with the Cambridgeshire dataset.

The class 3 showed the best accuracy among the classes (67%). Although not so high accuracy of class 3, other classes have higher accuracy than models obtained by the other two methods. This is the reason because the overall accuracy is little higher in the case of Gradient Boosting. The class 1 has 41% of accuracy, followed by the class 2 with 35%. We even have one image correctly classified in the class 5. The class 4 has 0% of accuracy.

Similarly like in Neural Networks model, we have a lot of images from classes 1 and 2 classified like class 3, and opposite, images from class 3 classified like classes 1 or 2.

In this case, situation is little better with values P and D than in Neural Networks based model. 34% of images are classified more in the class 3 than this class really contains. 16% less in the class 1 and 17% less in the class 2 than they really contain. In absolute numbers, only high value is in class 3, which is 55 images classified more like this class than it really exists in this class.

CONFUSION MATRIX (GRADIENT BOOSTING)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**40**	16	41	0	0	97	41%
2	16	**40**	58	0	0	115	35%
3	22	41	**108**	0	1	162	67%
4	1	1	0	**0**	0	2	0%
5	3	0	10	0	**1**	14	7%
Total (X)	82	98	217	0	2	**189**	Av. 30%
CC/X	49%	41%	50%	N/A	50%	Av. CC/X = 48%	
P=X/N	84%	85%	134%	0%	14%	Σ|P-100|/C=50%	T. Im.:**400**
D=X-N	-15	-17	55	-2	-12	Σ|D| = 101	**Acc. 47.25%**

Table 26. Confusion matrix of Gradient Boosting predictive model built on Coimbra district dataset

MBR method gave predictive model with the best results on this dataset (Table 27), with 50% of the overall accuracy.

CONFUSION MATRIX (MBR)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**37**	26	34	0	0	97	38%
2	14	**44**	57	0	0	115	38%
3	16	36	**119**	0	1	162	73%
4	2	0	0	**0**	0	2	0%
5	2	1	11	0	**0**	14	0%
Total (X)	71	107	221	0	1	**200**	Av. 30%
CC/X	52%	41%	54%	N/A	0%	Av. CC/X = 37%	
P=X/N	73%	93%	136%	0%	7%	Σ|P-100|/C=53%	T. Im.:**400**
D=X-N	-26	-8	59	-2	-13	Σ|D| = 108	**Acc. 50%**

Table 27. Confusion matrix of MBR predictive model built on Coimbra district dataset

The highest accuracy in the class 3 (73%) and relatively high accuracy in classes 1 and 2 (38% each) provided that this model have the best performances, and the best overall accuracy also.

Like in the previous cases, we have high number of wrongly classified images - from the class 3 classified like class 1 or 2 and opposite. But, in this case, confusion between classes 2 and 3 is more present.

In this model case, the situation is similar like in two previous. Only the class 2 is better distinguished in case of the MBR - only 7% of images were less classified like the class 2 than this class really contains. This is better if comparing with 15% images less in the Gradient Boosting based model. Value P for the class 3 is still high, 136%. In absolute numbers, it means that the class 3 contains 59 classified images like the class 3 more than it really exists in this class.

In the Coimbra district dataset, the MBR method has built the best model, the same like in the case of the Cambridgeshire dataset. This can be an indicator that the MBR method is the most suited for building predictive models on this kind of data.

4. 2. 3. South Bačka district dataset results - the second round

South Bačka district testing set has 400 images. The most frequent is class 1 with 197 images. Surprisingly, the second class by the number of images is class 5 with 79 images. This is maybe not so surprising if we know that Danube River is present in this region, and that people love the life on the river. A lot of boats and floating rafts along the river bank is a good prerequisite for taking photos with standing point in class 5. Classes 2 and 3 are almost equal with the number of images, 60 and 58, respectively. Class 4 contains 6 images.

The model with the best results built with the Neural Networks method (Table 28) has the overall accuracy of 70,5%.

The accuracy is very high in this case because of two reasons. The first is author's familiarity with Serbian language and the second is very high distinction among the classes (except between classes 1 and 2).

The class 1 has the highest accuracy (89%), like is the case with all most frequent classes. The class 3 also has reasonably high accuracy (81%). In the class 4 we have 5 correctly classified images (out of 6), which is 83% of the accuracy. The class 5 has accuracy of 59%. It is good if we compare with the class 5 in the other datasets. Quite high number of the wrong classified images in the class 5 is present. Most of them are classified like the class 1.

At the end, the lowest accuracy has the class 2 (13%). Majority of images from this class are classified like the class 1. The reason for this is that agricultural fields surround towns and villages. So, like the case was before, tagging one class, but taking picture from the standing point in another is also present here (Estima, Fonte and Painho, 2014; Zielstra, 2013).

CONFUSION MATRIX (NEURAL NETWORKS)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**175**	1	4	1	16	197	89%
2	37	**8**	7	0	8	60	13%
3	5	1	**47**	0	5	58	81%
4	1	0	0	**5**	0	6	83%
5	30	1	1	0	**47**	79	59%
Total (X)	248	11	59	6	76	**282**	Av. 65%
CC/X	70%	73%	80%	83%	62%	Av. CC/X = 74%	
P=X/N	126%	18%	102%	100%	96%	Σ|P-100|/C=25%	T. Im.:**400**
D=X-N	51	-49	1	0	-3	Σ|D| = 104	**Acc. 70.5%**

Table 28. Confusion matrix of Neural Networks predictive model built on South Bačka district dataset

If we observe value P in this model, we can conclude it is quite satisfying in classes 3, 4 and 5. Little higher value is present in the class 1, and very low value in the class 2. In absolute numbers, we have 1, 0 and -3 images classified more or less in classes 3, 4 and 5, respectively, which is good. The class 1 has 51 images more classified like this class than it really contains, and the class 2 has 49 images less classified like this class than it really contains.

Gradient boosting gave the model with the best results (Table 29) with 66.25% of the overall accuracy. Cleaning of data didn't help much in this case.

The class 1 has 88% of the accuracy. The class 3 is the second with 71%, followed with the class 5 with 56% of the accuracy. In the class 4 we have 3 correctly classified images out of 6 (50%). The lowest accuracy is in the class 2, only 5%.

The situation is similar like in the Neural Networks model, just with the lower accuracy in all classes. The gradient boosting method showed quite good distinguishing between the classes in this dataset, which is not the case with previous ones. After we expelled the Decision Trees and the Regression methods from detailed study, we can say now that Gradient Boosting is on the third place if we compare our 5 methods.

CONFUSION MATRIX (GRADIENT BOOSTING)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**174**	3	2	0	18	197	88%
2	43	**3**	6	0	8	60	5%
3	11	3	**41**	0	3	58	71%
4	2	0	0	**3**	1	6	50%
5	34	1	0	0	**44**	79	56%
Total (X)	264	10	49	3	74	**265**	Av. 54%
CC/X	66%	30%	84%	100%	59%	Av. CC/X = 68%	
P=X/N	134%	17%	84%	50%	94%	Σ|P-100|/C=38%	T. Im.:**400**
D=X-N	67	-50	-9	-3	-5	Σ|D| = 134	**Acc. 66.25%**

Table 29. Confusion matrix of Gradient Boosting predictive model built on South Bačka district dataset

Not so great values of P in this model. That is reflected also on the overall accuracy. From analysing values P and D in the previous models, we can conclude that they are correlated with the overall accuracy. As the overall accuracy is bigger as value P is closer to 100% and value D is closer to 0.

The model with the best results was built by MBR method (Table 30). It has 71% of the overall accuracy.

CONFUSION MATRIX (MBR)							
PREDICTED → / CLC ↓	1	2	3	4	5	Total (N)	Class Acc.
1	**173**	3	3	1	17	197	88%
2	33	**14**	5	0	8	60	23%
3	5	5	**46**	0	2	58	79%
4	1	0	1	**4**	0	6	67%
5	31	1	0	0	**47**	79	59%
Total (X)	243	23	55	5	74	**284**	Av. 63%
CC/X	71%	61%	84%	80%	64%	Av. CC/X = 72%	
P=X/N	123%	38%	95%	83%	94%	Σ|P-100|/C=23%	T. Im.:**400**
D=X-N	46	-37	-3	-1	-5	Σ|D| = 92	**Acc. 71%**

Table 30. Confusion matrix of MBR predictive model built on South Bačka district dataset

The class 1 has the highest accuracy (88%), which could be predicted (based on the previous cases) because this is the most frequent class. The class 3 is the second with 79% of the accuracy, while the class 5 has 59%. In the class 4, we have 4 correctly classified images out

of 6. One of the things that make this method the best performance method can be seen in the class 2, which is 23% of the accuracy, compared with 13% in the case of the Neural Network built model and 5% in the case of the Gradient Boosting built model.

This method managed to raise the class accuracy of the class 2 on 23%, comparing with 13% in the Neural Networks model. Other classes are quite similar like in the Neural Networks method model, maybe even little worse, but distinguishing the class 2 from the class 1 is significantly better, and, in this case, it makes the overall accuracy better in the MBR method built model.

This model has similar values of P and D like the Neural Networks model. The difference is that in this case values are better in classes 1 and 2, which are also more frequent, but values are little worse in classes 3, 4 and 5. These classes are not so frequent (except class 5), so, the overall accuracy is little better in the case of this model.

The MBR method built the model with the best performances in this dataset also. So, we can conclude that in the all 3 datasets, the MBR method based models showed the best results - performances and overall accuracy. The Neural Networks, like the second place method, is also worth mentioning because it is closely following MBR method with its performances and the overall accuracy.

4. 3. Overall Discussion

Already with analysing data from the first round of the data cleaning, it was possible to notice some general characteristics of the methods used. All methods have characteristics of high accuracy of the most frequent class. Neural Networks has high accurate results of the most frequent class, but low of the other classes. Gradient Boosting showed even higher accuracy of the most frequent class, but lower accuracy of other classes than Neural Networks. MBR didn't show the highest accuracy of the most frequent class, but it showed much higher accuracy of the other classes and it had the highest overall accuracy.

Decision Trees and Regression were showing similar behaviour. They had extremely high accuracy of the most frequent class, but extremely low accuracy of other classes (especially Decision Trees). These two methods had quite high overall accuracy, but their performances were very poor. That was the reason for excluding these two methods from further research. After researching more, it is noticed that Gradient Boosting also has some similar characteristics, which is not a surprise because it is based on Decision Trees.

As Cerrito (2009) discussed in her paper - is the best model this one with the highest overall accuracy or this which is able to distinguish low frequent classes from high frequent classes? Her answer was that appropriate method should be able to equally distinguish all classes. We agree with her. This is the reason why we chose MBR method like the best suited method for this type of data, which is text. Not only this, but MBR method had the highest overall accuracy in the cases with all three study areas. However, it must be said that Neural Networks method performed slightly worse than MBR, so, it should inevitably be included in further research on this type of data.

Overall accuracy of over 70% is significant result. Interpretation of this result could be that extracting land use information from Panoramio photo tags is possible, although it needs further research and improving results in order to become reliable source of information.

One significant issue was noticed in this study. It is standing point of photographer. The results are not higher mainly because of this appearance. Standing on the land and taking photo of the ocean or taking photo of a town from a distance are examples of this phenomenon (Estima, Fonte and Painho, 2014). Also, wrong geotagging and low positional accuracy of devices are causing that image is located on wrong land use type. These problems are evident. It would be necessary trying to solve them in further research.

At the end, the importance of data cleaning showed up like very big issue if we consider very bad results with the Coimbra district datasets comparing with other two areas. This is due to author's unfamiliarity with Portuguese language.

CHAPTER FIVE
5. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

5. 1. Conclusions

Folksonomy doesn't have strict structure as taxonomy has. Other factors are also causing that UGGC is not so usable in this case, like people's negligence or wrong geottaging. But potentials of UGGC sources are enormous. That is the reason why trying and succeeding in obtaining useful information from UGGC could be very useful and important (Sui, 2011).

We managed to build predictive models with more than 70% of accuracy. Such result doesn't mean that we provided total proof that extracting useful information (land cover) from people's expressions (tags) is possible, but this can be considered as a contribution to this problem. 70% of accuracy means that we proved the existence of some connections between photo tags and the land cover class on which a certain photo is taken.

Low accuracy in case of Coimbra district dataset is a proof that data cleaning process is a very important step in extracting information. This is due to the author's unfamiliarity with Portuguese language. The better accuracy obtained after the second round of cleaning is also a proof of the importance of the data cleaning step.

Another important issue is related with photos taken far from what they are showing, as already reported by (Estima, Fonte and Painho, 2014). Standing on one land cover class and taking photo and tagging other was noticed during the process of data cleaning. This might be the cause of almost 30% of error in the best models, as well as people's negligence in tagging photos (Sui, 2011).

In terms of algorithms, the MBR method showed the best results in building a predictive model with this type of data (text). Although, Neural Networks did not have really bad results, MBR expressed itself like the best method in this case.

We can conclude that extracting land use information from Panoramio photo tags is possible, although improving the accuracy is very important to make such sources of Geographic Information more reliable. We believe that an accuracy of over 70% proves the potential of these types of approaches. The MBR method showed the best results, but the other methods need to be extensively explored in different situations and areas to verify if we can generalise such conclusion or if this method performs better only in this particular case.

5. 2. Future Research Directions

For future work we can recommend several improvements:

- Improving the process of data cleaning, especially for automatic approaches
- Combining more sources of UGGC;
- Feeding models with more data to try to improve accuracy and eventually extract more information;
- Excluding images which are close to the borders of different land cover classes.

This study can be encouraging for future research on this kind of problems.

BIBLIOGRAPHY

1. AGGARWAL, C. C., and ZHAI, C. (2012) *Mining text data*. Springer.

2. CERRITO, P. B. (2009) *Predictive Modeling in Enterprise Miner Versus Regression.*

3. CIHLAR, J., and JANSEN, L. (2001) *From Land Cover to Land Use: A Methodology for Efficient Land Use Mapping over Large Areas.* The Professional Geographer, 53(2), 275–289. doi:10.1080/003301 24.2001.9628460

4. DAYHOFF, J. E., and DELEO, J. M. (2001) *Artificial neural networks.* Cancer 91.S8: 1615-1635.

5. ELWOOD, S., GOODCHILD, M. F., & SUI, D. Z. (2012) *Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice.* Annals of the Association of American Geographers, 102(3), 571–590. doi:10.1080/0004 5608.2011.595657

6. EGENHOFER, M. J., and MARK, D. M. (1995) *Naive geography*. Springer Berlin Heidelberg

7. ESTIMA, J. (2012) *Using Volunteered Geographic Information to help Land Use/Land Cover mapping.*

8. ESTIMA, J., FONTE, C. and PAINHO, M. (2014) *Comparative study of Land Use/Cover classification using Flickr photos, satellite imagery and Corine Land Cover database.* In: Huerta J, Schade S, Granell C (eds) Connecting a Digital Europe Through Location and Place. Proceedings of the 17th AGILE International Conference on Geographic Information Science. ISBN: 978-3-319-03611-3

9. ESTIMA, J. and PAINHO, M. (2013) *Exploratory analysis of OpenStreetMap for land use classification.* In: Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '13. ACM Press, pp 39-46. doi:10.1145/2534732.2534734

10. ESTIMA, J. and PAINHO, M. (2013) *Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover.* In: Murgante B, Misra S, Carlini M, Torre CM, Nguyen HQ, Taniar D, Gervasi O (eds) ICCSA 2013: Computational Science and Its Applications. The 13th

International Conference on Computational Science and Its Applications, Ho Chi Minh City, Vietnam, 24-27 June 2013 . Lecture Notes in Computer Science vol 7974. Springer, Heidelberg, pp 205-220. doi:10.1007/978-3-642-39649-6_15

11. ESTIMA, J. and PAINHO, M. (2014) *Photo Based Volunteered Geographic Information Initiatives: A Comparative Study of their Suitability for Helping Quality Control of Corine Land Cover*. In-ternational Journal of Agricultural and Environmental Information Systems 5(3): 75-92. doi: 10.4018/ijaeis.2014070105

12. ESTIMA, J. and PAINHO, M. (2015) *Investigating the Potential of OpenStreetMap for Land Use/Land Cover Production: A Case Study for Continental Portugal.* In: Jokar Arsanjani J, Zipf A, Mooney P, Helbich M, OpenStreetMap in GIScience: experiences, research, applications. ISBN:978-3-319-14279-1, PP. pending, Springer Press.

13. European Environment Agency (EEA). (2007) *CLC 2006 technical guidelines.* EEA Technical Report No 17/2007

14. FRIEDMAN, J. H. (2001) *Greedy function approximation: a gradient boosting machine.* Annals of Statistics: 1189-1232.

15. FRIEDMAN, J. H. (2002) *Stochastic gradient boosting.* Computational Statistics & Data Analysis 38.4: 367-378.

16. FOODY, G. M. (2002) *Status of land cover classification accuracy assessment.* Remote Sensing of Environment, 80(1), 185–201.

17. GOODCHILD, M. (2007) *Citizens as sensors: The world of volunteered geography.* GeoJournal, 69(4), 211–221. doi:10.1007/s10708-007-9111-y

18. GOODCHILD, M. (2008) *Assertion and authority: the science of user-generated geographic content.*

19. GOODCHILD, M. (2008) *Commentary: Whither VGI?* GeoJournal, 72(3-4), 239–244. doi:10.1007/s10708- 008-9190-4

20. GOODCHILD, M. and GLENNON, J. A. (2010) *Crowdsourcing geographic information for disaster response: A research frontier.* International Journal of Digital Earth, 3(3), 231–241. doi:10.1080/17538941003759255

21. HILL, L. L. (2000) *Core elements of digital gazetteers: placenames, categories, and footprints.* Research and advanced technology for digital libraries. Springer Berlin Heidelberg, 280-290.

22. HOLLENSTEIN, L. and PURVES, R. (2014) *Exploring place through user-generated content: Using Flickr tags to describe city cores.* Journal of Spatial Information Science 1: 21-48.

23.http://databases.about.com/od/datamining/g/regression.htm (Accessed: December, 2014)

24.http://en.wikipedia.org/wiki/Cambridgeshire (Accessed: December, 2014)

25.http://en.wikipedia.org/wiki/Coimbra_District (Accessed: December, 2014)

26.http://en.wikipedia.org/wiki/South_Ba%C4%8Dka_District(Accessed: December, 2014)

27.http://gis.stackexchange.com/questions/62715/corine-land-cover-2000coordinate-reference-system (Accessed: December, 2014)

28.http://land.copernicus.eu/pan-european/corine-land-cover (Accessed: December, 2014)

29.http://www.eea.europa.eu/data-and-maps/data/clc-2006-vector-data-version-3 (Accessed: September, 2014)

30.http://www.panoramio.com/ (Accessed: September, 2014)

31.http://www.panoramio.com/help/acceptance_policy (Accessed: December, 2014)

32.http://support.sas.com/documentation/cdl/en/emgsj/62040/HTML/default/viewer.hm#a 003307717.htm (Accessed: December, 2014)

33.http://support.sas.com/documentation/cdl/en/tmgs/62416/HTML/default/viewer.htm#n1gpxb3f6uwsu8n149yftcyonlql.htm (Accessed: December, 2014)

34.http://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf(Accessed: December, 2014)

35.http://support.sas.com/publishing/pubcat/chaps/57587.pdf (Accessed: December, 2014)

36.http://www.pcmag.com/encyclopedia/term/50809/sas-system (Accessed: December, 2014)

37. HUDSON-SMITH, A., BATTY, M., CROOKS, A. and MILTON, R. (2009) *Mapping for the Masses: Accessing Web 2.0 Through Crowdsourcing.* Social Science Computer Review, 27(4), 524–538. doi:10.1177/0894439309332299

38. HUGHES, M., O'CONNOR, N. E. and JONES, G. JF. (2012) *A machine learning approach to determining tag relevance in geotagged Flickr imagery.* Image Analysis

for Multimedia Interactive Services (WIAMIS), 13th International Workshop on. IEEE.

39. KENNEDY, L. et al. (2007) *How flickr helps us make sense of the world: context and content in community-contributed media collections.* Proceedings of the 15th international conference on Multimedia. ACM.

40. KISILEVICH, S. et al. (2010) *Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections.* Information Visualisation (IV), 2010 14th International Conference. IEEE.

41. LEUNG, D. and NEWSAM, S. (2010) *Proximate sensing: Inferring what-is-where from georeferenced photo collections.* Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE.

42. LEUNG, D. and NEWSAM, S. (2012) *Exploring geotagged images for land-use classification.* Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia. ACM.

43. MAINGI, J. K., KEPNER, S. E. and EDMONDS, W. G. (2002) *Accuracy Assessment of 1992 Landsat-MSS Derived Land Cover for the Upper San Pedro Watershed(US/Mexico).* 71 Sponsored by Environmental Protection Agency, Las Vegas, NV. National Exposure Research Lab.

44. MANYIKA, J. et al. (2011) *Big data: The next frontier for innovation, competition, and productivity.*

45. MARLOW, C. et al. (2006) *HT06, tagging paper, taxonomy, Flickr, academic article, to read.* Proceedings of the seventeenth conference on Hypertext and hypermedia. ACM.

46. MASAND, B., LINOFF, G. and WALTZ, D. (1992) *Classifying news stories using memory based reasoning.* Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.

47. NEWSAM, S. (2010) *Crowdsourcing what is where: community-contributed photos as volunteered geographic information.* IEEE Computer Society 17(4):36–45

48. Peters, I. (2009) *Folksonomies: Indexing and retrieval in Web 2.0*. Vol. 1. Walter de Gruyter.

49. SENSEMAN, G. M., BAGLEY, C. F and TWEDDALE, S. A. (1995) *Accuracy Assessment of the Discrete Classification of Remotely-Sensed Digital Data for Landcover Mapping.* DTIC Document.

50. SPECIA, L. and MOTTA, E. (2007) *Integrating folksonomies with the semantic web.* The semantic web: research and applications. Springer Berlin Heidelberg. 624-639.

51. SPYRATOS, S. and LUTZ, M. (2014) *Characteristics of Citizen-contributed Geographic Information.*

52. SUI, D. and GOODCHILD, M. (2011) *The convergence of GIS and social media: challenges for GIScience.* International Journal of Geographical Information Science 25.11 : 1737-1748.

53. TURNER, A. J. (2006) *Introduction to Neogeography* (M. O'Reilly, Ed.). Sebastopol, CA.

54. WANG, J., KORAYEM, M. and CRANDALL, D. J. (2013) *Observing the natural world with Flickr.* Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. IEEE.

55. ZIELSTRA, D. and HOCHMAIR, H. H. (2013) *Positional accuracy analysis of Flickr and Panoramio images for selected world regions.* Journal of Spatial Science 58.2 : 251-273.

ANNEXES

Annex A

Tag	Num	Tag	Num	Tag	Num	Tag	Num
"railways"	168	"Abbey"	32	"Abington"	28	"aerial"	12
"aircraft"	25	"Berdfordshire"	55	"Anglesey"	23	"Ashwell"	20
"Astwick"	35	"Audley End"	33	"autumn"	14	"balloon"	16
"Bedford"	69	"best"	55	"blue"	7	"British Isles"	37
"Burghley"	29	"Burwell"	15	"business"	3	"Cambridge"	338
"clouds"	12	"Duxford"	32	"Ely"	156	"fire"	1
"Huntingdonshire"	178	"Wisbech"	53	"canabis"	1	"cars"	27
"Land Rover"	19	Symbols	183	"Norfolk"	165	"Suffolk"	138
"Essex"	59	"Hertfordshire"	135	"Peterborough"	54	"Royston"	52
"Stamford"	26	"favourite"	7	"Hinxworth"	63	"Huntingdon"	139
"I love England"	5	"Kempston"	33	"Langford"	27	"Lincolnshire"	37
"London"	5	"macro"	5	"morning"	2	"Oxford"	1
"people"	6	"places"	4	"Potton"	7	"Saffron Walden"	139
"Shefford"	59	"St. Neots"	148	"storm"	3	"Stotfold"	33
"sunset"	3	"Swansey"	23	"telephone"	1	"Thorney"	33
"traditional"	3	"trains"	25	"travel"	19	"Whittlesey"	39
"Wilburton"	27	"sky"	6	"sunrise"	3	"snow"	3
"nikon"	1	"Milton"	12	"reino unido"	1	"Orsolya"	2
"butterfly"	1	"summer"	3	"spring"	2	"merged"	1
"Google Earth"	25	"winter"	3	"night"	3	"vehicles"	7
"walking"	3	"tio alberto"	1	"moon"	1	"CL sites"	17
"Jubilee 2012"	1	"Kimbolton"	23	"Knapwell"	28	"Magdalena"	1
"Molesworth"	3	"signs"	16	"Stilton"	17	Total:	3,763 (53.66%)

Table A. Removed photos' tags in the Cambridgeshire dataset in the second round of data cleaning

Annex B

Tag	Num	Tag	Num	Tag	Num	Tag	Num
Symbols	648	"A14"	58	"A17"	36	"Figueira da Foz"	498
"aereal"	15	"Aigra Nova"	55	"Aldeia de Nogueira"	47	"Aldeia de Soito"	38
"Aldeia de Xisto"	3	"Aldeia do Tojo"	2	"Aldeia Velha"	15	"Aldeias"	13
"Alfarelos"	3	"Algarve"	1	"Alhadas"	32	"Almedina"	78
"Alminhas"	89	"Alvares"	101	"Alvorge"	82	"amigos"	8
"Anadia"	15	"best"	142	"Leiria"	158	"Arganil"	275
"Arredores da Pena"	42	"arte"	57	"automoveis"	7	"Aveiro"	3
"Baixo Mondego"	28	"Beira"	135	"black & white"	12	"Bobadela"	139
"bom sucesso"	3	"bom viagem"	1	#Borda do Campo"	23	"Bordeiro"	11
"Bussaco"	13	"Cabanas de Viriato"	29	"Cabo do Mondego"	37	"canon"	3
"Carregal do Sal"	48	"carro"	25	"Carvarhal do Sapo"	12	"Casconho"	17
"Castanhera de Pera"	16	"Coentral"	43	"Coimbra"	682	"Coimbra district"	395
"Colmeal"	38	"Concelho Arganil"	95	"Concelho de Penela"	86	"Condeixa"	31
"Condeix-a-Nova"	71	"Conimbriga"	30	"Couchel"	28	"Curia"	5
"curiosas"	9	"descobertas"	3	"diversos"	5	"energia"	15
"Espinhal"	48	"Esquio"	39	"Felgueira Velha"	73	"flores"	32
"fotografias"	3	"Freixo"	14	"Funnyportugal"	3	"Gois"	117
"Google Earth"	23	"Granja de Ulmeiro"	15	"Janeiro de Cima"	29	"Lagares da Beira"	32
"Lara em Fartosa"	43	"localidades"	13	"Lomba do Bargo"	31	"look around"	1
"Lugar da Picha"	7	"Luso"	194	"Mega Fundeira"	78	"meu avo"	1
"Milrico"	22	"Mira"	37	"Miranda do Corvo"	69	"Moita da Serra"	38
"Montemor-o-Velho"	43	"my tags"	10	"noite"	1	"nova technologia"	1
"Obras"	85	"Oleiros"	103	"outono"	4	"outras"	6
"paisagens"	8	"Pampilhosa da Serra"	27	"Pe de Esquio"	129	"Pena"	158
"Penacova"	176	"Penela"	122	"pessegueiro"	52	"pets"	1
"Pinheirinho"	85	"Pomares"	35	"Lapa do Lobo"	29	"Quinta de Belide"	48
"Quinta do Fim"	1	"Quinta do Prazo"	3	"quintal da minha tia"	1	"S. Martinho do Bispo"	43
"Santa Luzia"	58	"Sao Pedro de Alva"	92	"Sarzedas do Vasco"	83	"Seixo de Beira"	46
"www.fotodoze.com"	3	"Luso Bussaco"	23	Total:		7,047 (69.81%)	

Table B. Removed photos' tags in the Coimbra district dataset in the second round of data cleaning